Master's thesis

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

Harald Aarskog
Johannes Lindstad

An integrated approach to tactical resource and admission planning in a cancer clinic

Master's thesis in Industrial Economics and Technology Management

Supervisor: Anders N. Gullhav, Bjørn Nygreen

June 2020

NTNU
Norwegian University of
Science and Technology

Harald Aarskog
Johannes Lindstad

An integrated approach to tactical resource and admission planning in a cancer clinic

Master's thesis in Industrial Economics and Technology Management
Supervisor: Anders N. Gullhav, Bjørn Nygreen
June 2020

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 thesis marks the end of our Master's degree in Industrial Economics and Technology Management at The Norwegian University of Science and Technology (NTNU). The work is done in collaboration with Oslo University Hospital and is intended to serve as a decision support tool for the tactical patient admission and resource planning for cancer patients in a multi-disciplinary care system. The thesis builds on a report submitted in the course TIØ4500 Managerial Economics and Operations Research, Specialization Project (Aarskog and Lindstad, 2019).

We would like to thank our supervisor Anders N. Gullhav and our co-supervisor Bjørn Nygreen for their valuable feedback during the project period. We are also very thankful for feedback and information provided by Erik Rokkones, Ingrid Kristine Small Hanto and Per Magnus Mæhle at Oslo University Hospital.

This thesis was written during the global pandemic of COVID-19. Within a few weeks, the virus went from something we chatted about during lunch hour, to something that heavily affected our working life. We quickly had to adapt from physical meetings with our supervisors and with each other, to home office, Skype conversations and e-mailing. We thank NTNU for giving us support and follow-up during this extraordinary period and the ability to finish the thesis on campus.

Trondheim, June 10th, 2020

Harald Aarskog Johannes Lindstad

Harald Aarskog Johannes Lindstad

Summary

Hospitals and health care institutions worldwide are facing a challenging situation where they have to increase the quality of their services, while reducing costs. This thesis presents an application of operations research for cancer patients residing to the Department of Gynecological Cancer at Oslo University Hospital, where analytical methods and simulation are combined for efficient patient admission and resource allocation for cancer patients.

We assess the problem of tactical patient admission and resource planning for patients enrolled in a standardized care process in a multi-disciplinary care system. The problem covers multiple resources and multiple standardized care processes and its objective is to minimize the waiting time for patients enrolled in a care process. By keeping track of queues linked to each hospital activity in each care process, resources are allocated to the correct activity at the right time, ensuring satisfying and timely care. When the problem is solved, a tactical resource and patient admission schedule is generated. Decisions must adhere to pivotal restrictions regarding time, capacity, sequencing and queues.

We present a Mixed Integer Programming (MIP) model that aims at solving the resource and admission planning problem presented above. The problem is solved for a planning period that spans over a few weeks using a commercial MIP solver. The work is intended to serve as a decision support tool for hospital management facing planning problems with similar characteristics as the one presented in this thesis. The work may assist in the planning process of patients enrolled in a standardized care process by suggesting alternative schedules for the patients and resources, and by automating work that to a large distinct today is carried out manually.

Over the course of one such planning period, the dynamics of the system are not revealed. To evaluate and appraise the results of the optimization model and dynamics of the system over multiple planning periods, we suggest a scheduling framework that takes on a rolling horizon approach. The sole purpose of the scheduling framework is to provide an environment that imitates the reality of the hospital where the solutions from the optimization model may be tested. The hospital does not conduct any simulations in their implementation. In the scheduling framework, the optimization model is solved for a planning period. Then, parts of this planning period is simulated. After the simulation, the optimization model is run again for a new planning period, commencing from the day the simulation left off.

Our work offers an alternative way of allocating resources and scheduling patients enrolled in a standardized care process. First, we are able to handle time limits for patients enrolled in a care process. Also, by combining optimization and simulation in the scheduling framework, we are able to first generate optimal or near optimal solutions for the optimization problem and evaluate the framework's performance in a stochastic environment using simulation. The scheduling framework is updated with new schedules for the coming planning period using the rolling horizon approach. For the rolling horizon approach, we have developed a method that reduces the number of adjustment from one schedule to the next in order to provide predictability for the hospital staff and the patients that are to be serviced.

We provide a case study inspired by the Department of Gynecological Cancer at Oslo University Hospital and simulate how our model performs. Using our work, it is found that 90 % of patients are able to start their treatment within the limits decided by The Norwegian Directorate of Health. Results also show that by increasing the length of the planning period, waiting times decrease. We also find that by decreasing the implementation period, that is running the optimization model more often, waiting times decrease.

This master's thesis succeeds in exploring how operations research methods may be utilized to schedule patients enrolled in a standardized care process. The optimization model succeeds in achieving the goal of minimizing waiting times and allocating multiple resources to patients with different cancer diagnoses, taking into account interdependence and order constraints. We are also able to validate the model and the system's performance under uncertainty in a real life and dynamic environment using simulation. In conclusion, the optimization model may serve as a decision support tool in a hospital planning process and the simulation model is used to evaluate the performance of the solutions from the optimization.

Sammendrag

Sykehus og helseinstitusjoner over hele verden står overfor en utfordrende situasjon hvor de både må øke kvaliteten på tjenestene sine og kutte kostnader. Denne masteroppgaven presenterer en anvendelse av operasjonsanalyse for kreftpasienter tilknyttet Avdeling for gynekologisk kreft ved Oslo Universitetssykehus. Vi kombinerer analytiske metoder og simulering for å effektivt planlegge og allokere ressurser for kreftpasienter.

Vi undersøker det taktiske planleggingsproblemet for å planlegge og allokere ressurser for kreftpasienter som tar del i et standardisert pasientforløp, eller pakkeforløp. Problemet inneholder flere ressurser og flere pakkeforløp, og dets objektiv er å minimere ventetiden for pasienter i pakkeforløp. Ved å overvåke køene til hver sykehusaktivitet i hvert pakkeforløp, allokerer vi riktig mengde ressurs til riktig tid, og forsikrer riktig pleie til riktig tid. Når problemet løses, genereres det taktiske planer for når pasienter skal gjennomgå ulike sykehusaktiviteter og hvilke ressurser som settes av til hvilke dager. Beslutninger må overholde restriksjoner som omhandler tid, kapasitet, rekkefølge og køer.

Vi presenterer en blandet heltallsmodell som sikter på å løse problemet introdusert ovenfor. Problemet løses for en planleggingsperiode som strekker seg over noen uker. Vi bruker en kommersiell programvare til å løse det blandede heltallsproblemet. Optimeringsmodellen er ment til å fungere som et verktøy som støtter beslutningstaking for planleggingsproblemer med samme karakteristikker som problemet vi presenterer. Oppgaven fungerer som en støtte i planleggingsarbeidet for pasienter i et standardisert pasientforløp ved å foreslå alternative timeplaner for pasienter og ressurser, og ved å automatisere arbeid som i dag kan være manuelt og tidkrevende.

Ved å løse optimeringsproblemet for én planleggingsperiode, blir ikke dynamikken til systemet fullt ut avdekket. For å evaluere og etterligne systemets dynamikk over flere planleggingsperioder, foreslår vi et planleggingsrammeverk med en rullerende horisont-tilnærming. I planleggingsrammeverket løses optimeringsmodellen først for én planleggingsperiode som typisk varer noen få uker. Løsningen av optimeringsmodellen etterfølges av en simulering av løsningen fra optimeringen. Etter simuleringen blir optimeringsproblemet løst på nytt for en ny planleggingsperiode som starter i den tidsperioden simuleringen sluttet. Simuleringen tilfører ikke noe til optimeringen og er ikke noe sykehuset gjennomfører. Simuleringen er heller å regne som et substitutt for faktisk implementering av modellen på sykehuset.

Vårt arbeid foreslår en alternativ tilnærming som allokerer ressurser og planlegger prosessen for pasienter som er innlemmet i et standardisert pasientforløp. For det første håndterer vi tidsfrister. Videre, ved å kombinere optimering og simulering i et planleggingsrammeverk, klarer vi å generere optimal eller nesten optimal løsning for probleminstanser av en viss størrelse og evaluere systemets prestasjon i et stokastisk miljø ved å simulere. Planleggingsrammeverket oppdateres med nye timeplaner for den kommende planleggingsperioden ved å ha en rullerende horisont-tilnærming til problemet. For den rullerende horisonten har vi utviklet en metode som reduserer antallet endringer fra en timeplan til den neste. Dette er gjort for å gi stabilitet og forutsigbarhet for ansatte på sykehuset og pasientene.

Vi presenterer også et eksempelstudie inspirert av Avdeling for gynekologisk kreft på Oslo Universitetssykehus og simulerer hvordan vår modell yter i denne settingen. Her finner vi at 90 % av alle pasienter i et pakkeforløp starter behandlingen sin innenfor tidsfrister gitt av Helsedirektoratet. Vi undersøker også effekten av å variere parametere i planleggingsrammeverket. Resultatene viser at ved å øke lengden på planleggingsperioden, går ventetiden ned. Vi finner også at ved å redusere lengden på implementeringsperioden, altså ved å løse optimeringsmodellen oftere, så minker ventetiden.

Denne masteroppgaven lykkes i å granske hvordan operasjonsplanlegging kan brukes til å timeplanfeste aktiviteter og allokere ressurser for pasienter i standardiserte pasientforløp. Optimeringsmodellen oppfyller målet om å minimere ventetider for pasienter og å allokere flere ressurser til pasienter med ulike diagnoser. Modellen takler også å ta høyde for gjensidige avhengigheter i systemet og rekkefølgekrav for aktiviteter. Vi validerer modellen og systemets ytelse under usikkerhet i et virkeligsnært og dynamisk miljø ved å bruke simulering. I tillegg klarer simuleringsmodellen å evaluere ytelsen til løsninger som optimeringsmodellen finner.

Table of Contents

List of Figures

List of Tables

Glossary

Activity	A component of a cancer care pathway.
Appointment	A planned consultation with a medical professional.
Arrival rate	Number of patients arriving per time unit.
Age-standardized rate	A procedure for adjusting rates, e.g. incidence rates, designed to minimize the effects of differences in age composition when comparing rates for different populations. Referred to as age-standardized (or age-adjusted) rates.
ASR	Age-standardized rate.
Care process	Sequence of patient activities.
Cancer care pathway	Sequence of activities performed in a cancer care process.
Cancer stage	Classification of severity of cancer based on the tumor and its effect on the patient. I is the least severe and IV is the most severe.
FIFO	First in, first out.
GP	General practitioner.
Idle time	Non-productive time.
MDT	Multi-disciplinary team.
MIP	Mixed Integer Programming.
Multi-disciplinary care	Care involving multiple interrelated appointment comprising physical resources and heath care personnel from different hospital units.
Outpatient clinic	Hospital unit specialized in treating patients with health problems that needs diagnosis or treatment, but are not requiring a bed or to be admitted for overnight care.
Precedence constraint	Specification of the sequence of appointments.
Rolling horizon	Planning approach where plans initially are generated for time periods 0 to N. In the next iteration, plans are updated for periods 1 to $N + 1$, and so on.
Shared resource	Resource that is not exclusively ”owned” by one hospital unit, but shared among a set of hospital units. Typical examples are scanners in a CT lab where each hospital unit is assigned a quota per planning period.
Waiting time	Time between the end of one activity and the beginning of the next in a patient's cancer care pathway.
Simulation period	The time periods that are simulated in a simulation study.
WHO	World Health Organization.

Source: Most definitions are based on Leeftink et al. (2018) and Cancer Registry of Norway (2019).

1 Introduction

A cancer diagnosis is shattering for the person in question and next of kin. Due to a larger and aging population, and the fact that risk of cancer grows by the age, the number of deaths from cancer is expected to increase in the years to come. This is happening, paradoxically, in a time where screening processes are more thorough than ever before and novel treatment methods are being introduced. The cost of Norwegian health care services was 10.4 % of Norway's GDP, that is an astounding 342 billion NOK, in 2017 (Statistics Norway, 2018). Gynecological cancer is one of the most prevalent types of cancer for Norwegian women, with approximately 1800 new incidences per year, according to the Cancer Registry of Norway (2020). The survival rate for most types of gynecological cancer is high if the cancer is discovered at an early stage. To even further increase the survival rates, decreasing patient waiting times for diagnostics and treatment is vital.

In order to deliver reliable health care services to an increasing population under tight capacity, hospital planners and management are forced to organize processes more efficiently. All patients that need a treatment, should be serviced as soon as possible. At the same time, some patients will need to be prioritized over others, e.g. cancer patients. Hospitals are often organized in organizational silos and planners may tend to focus on one planning area at the time, e.g. the outpatient clinic, diagnostic services or operating rooms. This myopic type of planning often leads to local sub-optimization, which in turn leads to bullwhip effects in the care chain and longer waiting times. In the treatment of cancer, resources across different departments must be coordinated in order to swiftly deliver a diagnosis, treatment and follow-up to each individual patient. The challenge is that hospital departments are not coordinating their activities and are unable to deliver adequate care for all cancer patients. Under today's planning regime, the guidelines on time limits in cancer care addressed by the Norwegian Directorate of Health are seldom complied to by Norwegian hospitals providing cancer care.

In this thesis we address a tactical planning problem inspired by a Norwegian hospital. The facility is specialized in care for patients diagnosed with gynecological cancer. The situation today is that there exists challenges in the coordination between hospital units, meaning, for example, that decisions made in the radiology department not necessarily adhere to the plans of the gynecologists. This asymmetry in decision-making and planning might often lead to inefficient employment of resources and implicitly longer waiting times for patients waiting to start their treatment. The problem is constrained by time, capacity, precedence and resource restrictions. Our work does also include heterogeneous patient groups with a pre-defined set of possible care pathways. Each care pathway consists of a set of stages from diagnosis to evaluation and control. Each stage in a cancer care pathway (CCP) consists of a set of activities (e.g. a CT scan, a gynecological examination). This thesis' focus is on the two first stages of the cancer care pathway, that is from diagnosis up until treatment. The purpose of this thesis is to develop a decision-support tool for tactical resource planning in a hospital. We aim at developing an optimization model that minimizes waiting times for gynecological cancer patients from the referral from the GP to they have finished their treatment, by facilitating coordination between hospital units and effective resource utilization while at the same time adhering to the imposed restrictions.

Coordinating multiple medical disciplines in order to provide better and more integrated health care services is becoming increasingly important. There are more patients than ever before, diagnoses are more complex and patients often have comorbidities. These tendencies all point towards a more integrated approach to planning and control of health care services, where multiple medical disciplines form a multi-disciplinary system and patient care is organized and planned in a collaborative manner in order to provide shorter waiting times. Many papers have recently pointed towards the negative effects linked to long waiting times in health care, in particular cancer care. Long waiting times may lead to increased risk of the cancer recurring (Chen et al., 2008), unnecessary emotional strain and anxiety (Risberg et al., 1996; Rutqvist, 2006; Mackillop, 2007) and possibly a worsened prognosis (Richards, 2009; Hansen et al., 2011; Sorensen et al., 2014). Despite this, the 2019-proportions of gynecological cancer patients that were part of a standardized cancer care pathway and started their treatment within the recommended time limits from The Norwegian Directorate of Health, were 65 %, 60 % and 56 % for ovarian, uterine and cervical cancer, respectively (Norwegian Patient Registry, 2020). These numbers are below The Norwegian Directorate of Health's ambition of 70 %.

The motivation for this thesis is two-parted. The first perspective is from the patient's point of view. As underlined above, the negative impacts of waiting on each patient's health, both physically and mentally, are indisputable. The conclusions found in the literature are obvious - the shorter the time from the first appearance of cancer symptoms to diagnosis to start of treatment, the better the overall prognosis for the patients. It is also the patients that pay for the health care services, either through taxes or direct payments. Thus, providing health care may be viewed upon as delivering products to a customer that has paid for a service, and not for a queue ticket for that same service. Secondly, it is in the interest of the society as a collective to have healthy members and provide adequate health care. As long as demand exceeds supply, hospital managers are obliged to reasonably prioritize and optimize the health care system in a manner that ensures efficient utilization of resources

To solve this problem we adopt a system-wide approach and look past the traditional silo mentality that may be present in hospitals. This is done to better integrate hospital functions and facilitate coordination between different departments involved in patient care. We also evaluate the impact a decision made in one hospital department has on other hospital departments to prevent biased and sub-optimal decision-making. Both analytical methods and simulations may be utilized to solve scheduling and planning problems related to multi-disciplinary systems in health care. Analytical methods may provide an optimal or close to optimal solution, but lack the ability to readily solve more complex problems in a dynamic environment (Hulshof et al., 2013). Contrarily, simulations may be applied to model more complex systems, but they do not guarantee finding the optimal solution or a satisfying solution at all. The approach that is taken in this thesis is to develop an analytical Mixed Integer Programming (MIP) model that develops resource and admission plans for patient groups that require multi-disciplinary care and to test and evaluate this model based on extensive simulations of it. The simulation is also used to analyze how the dynamics of the problem are preserved when the optimization problem is solved over multiple planning periods.

Although literature on multi-disciplinary planning in a health care setting already exists, there is a void in the literature in terms of scheduling of multiple patient groups and multiple resources simultaneously. One of the contributions of this thesis is that it combines an analytical approach with simulation. The analytical optimization model is responsible for finding an optimal or near optimal solution to the problem. The role of the simulation is to evaluate how the solution of the optimization model performs in a realistic setting, where planning is done iteratively over multiple planning periods and uncertainty is taken into account. Another contribution is the weighted sum objective function, where objective function weights are assigned to each queue in a manner that minimizes a weighted sum of patient waiting times. This objective function formulation is important for at least two reasons. First, system interdependence between hospital units, that is, having a system-wide approach, may be administered by adjusting the objective function weights. Secondly, it enables an efficient management of patient flow, from intake to evaluation and control, by enabling hospital management to implicitly bring in prioritization rules for some patients requiring precedence. The thesis also contributes to existing literature by consolidating cancer care pathway-specific time limits and flow shop scheduling.

This paper has been organized as follows: Chapter 2 presents the background for the multi-disciplinary planning problem in a health care setting. Chapter 3 provides relevant literature for the scientific setting. Chapter 4 describes the problem. Chapter 5 presents a mathematical model describing the problem. Chapter 6 introduces the simulation model. Chapter 7 provides the architecture of a scheduling framework that combines the optimization model and the simulation to assess the dynamics of the system over multiple planning periods. Chapter 8 provides input data for the optimization and simulation models and Chapter 9 provides a computational study. Chapter 10 states the concluding remarks of the work. Lastly, Chapter 11 provides an outlook on future research topics.

2 Background

As patients get more complex diseases and comorbidities, a holistic approach to care is necessary. It is seldom the case that one doctor or a single hospital department is able to provide all the required care for a patient. Therefore, it is an apparent need to leave the nearsighted Taylorism behind, where the focus is on optimization of one system component at the time, and rather take on an integrated approach to multi-disciplinary planning and coordination in health care (Vanberkel et al., 2009). We operate with the same definition of a multi-disciplinary care system as Leeftink et al. (2018): *A multi-disciplinary care system is a care system in which multiple interrelated appointments per patient are scheduled, where health care professionals from various facilities or with different skills are involved.*

In this chapter relevant background material for the thesis is presented. The material presented is included to better understand what an integrated approach to multi-disciplinary planning and control in health care is and its importance. We divide health care into three geographical levels: global, national and hospital (Figure 2).

In order to contextualize the thesis, Section 2.1 introduces current and future global challenges in health care and discusses how these may be undertaken. Section 2.2 targets cancer care in Norway, corresponding to the second planning level in Figure 2. In Section 2.3 we provide a brief overview of gynecological cancer care at Oslo University Hospital today, which corresponds to the inner circle of Figure 2.

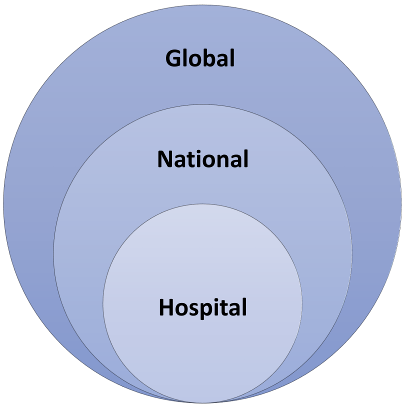


Figure 2: Health care planning at three different geographical levels. The main focus of this thesis is on the hospital level.

2.1 Global challenges in health care and the importance of an integrated approach to multi-disciplinary planning and control

In this section, a brief overview of global trends and challenges in health care is presented. This section aims at putting the thesis' content into a global context and to motivate the importance of an integrated approach to multi-disciplinary planning and control in health care.

The world's population grows steadily, and it has experienced a doubling of its population from 1973 to 2020 (United Nations, 2019a). This rapid growth, which is visualized in Figure 3, is expected to continue in the coming years, but at a slower pace. Within the next 40 years, the world's population is predicted to hit 10 billion (United Nations, 2019b). The population is not only growing in numbers, it is also aging at a fast pace. The UN predicts that by the year 2060, the percentage of the world's population over the age of 60 will double compared to the current situation. In addition, more people are lifted out of poverty, with the consequence that a greater proportion of the world's population will demand access to more sophisticated health care services, for example cancer care. This shift, facilitated by increased consumerism and a raise of expectations to new medical technology, points towards an even higher pressure on existing health care infrastructure, and an urgent need to develop smart approaches to planning and control of health care (Hurst, 2000).

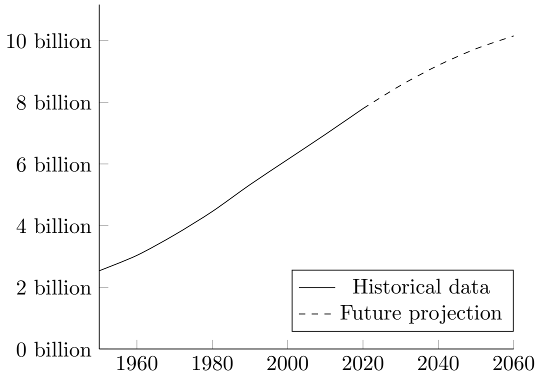


Figure 3: World population from 1960 to 2020 and future projection.

Source: United Nations (2019a,b)

Modern health care, and cancer care in particular, requires multiple hospital departments and other health caring units to be involved in the patient care process. A typical patient care process may start with a visit to the general practitioner (GP), followed by a stay at the hospital where the patient is diagnosed, a treatment stage and re-visits to the hospital for follow-ups and to the physiotherapist for rehabilitation. Planning and control of each of these activities can not be done in isolation from the others, since the outcome of one activity affects later actions. There are evidently interrelations between each of the system components, and this must also be reflected in the planning. Such interrelations involve: which unit performs which activity on a patient, when is a patient granted capacity in a health caring unit, in what order should the patient receive the care she is demanding and which factors decide how patients are prioritized?

Hospitals may often be rigorously organized in strictly separated units that suffer under silo mentality. The silo mentality often arises due to divergent goal-setting and reward schemes for different hospital units, lack of information sharing, relative physical distance between hospital units or medical specialty. To exemplify the divergent reward schemes, we introduce a short example: one hospital unit may seek to treat patients of type A because this grants them the highest compensation, while another interdependent hospital unit may want to maximize the number of treated patients of type B, based on the reward it receives. This asymmetry in behavior causes the system to act incoherently. The decision on which patients to treat has in this case become a political and economical issue, and not a medical one. The individual hospital units are not to be blamed for this behavior. The silo mentality is a structural problem that can only be fixed by actively motivating closer integration between the hospital units.

One major obstacle that makes multi-disciplinary planning demanding is that it restricts the room of maneuverability by introducing many constraints that would not be present if each unit planned their activities separately. This becomes evident, inter alia, in the allocation of shared hospital resources. Shared hospital resources are not exclusively owned by one hospital unit, e.g. a gynecological cancer department, but are rather shared among a set of hospital units. CT and MRI scanners are typical examples of shared resources in a hospital. The hospital units that share the resource may then book appointments at the shared resource up to a certain maximum limit. If they reach their limit, they may be granted more capacity, but then again, after negotiations and at the expense of other hospital units. With a fully integrated approach to planning between the involved hospital units, the capacity of the shared resources would be allocated to the different hospital units by a centralized system. This system should be based on pre-determined standardized rules that consistently would ensure maximum utilization and efficiency of the system as a whole.

At the same time, these are the same challenges, if handled properly, that can make multi-disciplinary planning so rewarding. By incorporating multi-disciplinary planning and control, traditional silo mentality is destroyed in favor of a collaborative way of working, where different health caring units involved in patient care work together as one collective unit. If the various hospital units involved in care of a type of patients, for example cancer patients, plan their activities in a coordinated manner, predictability would go up and waiting times for each patient would go down. The significance of the latter was highlighted on a patient level in Chapter 1. Non-coordinated use of shared hospital resources takes a toll on the degree of utilization of the relevant resources. Thus, with an integrated approach to multi-disciplinary planning, the hospitals would be enabled to treat more patients without it going at the expense of the quality of care.

It is noted that this thesis does not assess the monetary costs of implementation, nor the potential savings of planning hospital activities in a multi-disciplinary manner, but rather analyze the likely benefits in terms of waiting times and patient flow. However, according to industrial supply chain management theory, which is presented briefly in Section 3.4, it is expected that overall health care expenditures were to come down if multi-disciplinary planning became the norm in the health care sector. This is due to the decrease of the aforementioned bullwhip effect, that is, removing some of the variability in demand through the care chain by promoting collaboration and information sharing between the health caring units involved in care.

2.2 Cancer care in Norway

In this section, background material on cancer care in Norway is presented. A brief introduction to historical trends in cancer occurrences and treatment is presented, followed by a description of the standardized cancer care pathways in Norway. Finally, strengths and weaknesses of standardized cancer care pathways are briefly discussed.

One in three Nowegian will experience to be diagnosed with cancer at one point in life, according to the most recent annual report on cancer in Norway (Cancer Registry of Norway, 2019). As shown in the left plot of Figure 4, there is an increasing, but slower than before, trend of cancer occurrences in Norway for both sexes. Cancer incidences are measured by the age-standardized rate (ASR). This rate gives the number of yearly incidences per 100 000 person in the reference population. The reference population used in this report is the Norwegian mid-year population in 2014. It is noted that if one were to use another reference, e.g. The World Standard Population (Ahmad et al., 2001), the rates would differ. Therefore, the ASRs presented in this thesis are not comparable to metrics found in papers using other reference populations.

The main driver for the overall increasing trend in occurrences are owed to the relative aging of the population, and only a small share of the increase can be ascribed an actual increase in risk of cancer (Tretlie, 2016). Norwegian health authorities are pointing at efficient screening programs for cancer types only prevalent in women (e.g. breast cancer and gynecological cancers) to explain the more aggressive trend for women than for men (Johannesen, 2014; Helse- og omsorgsdepartementet, 2013). Risk of cancer may be linked to lifestyle choices, and according to WHO (2007) between 30 and 50 % of all cancer cases could be prevented if people smoked less tobacco, consumed less alcohol, ate more healthy food and polluted the air less.

The right plot in Figure 4 is showing a positive trend in survival rates for cancer patients in Norway. Men saw a more rapid increase than women in 5-year relative survival rates in the period illustrated. This is possibly explained by medical breakthroughs in cancers only prevalent in men. For example, new medical technology has made it possible to reach 5-year survival rates of almost 100 % for prostate and testis cancer.

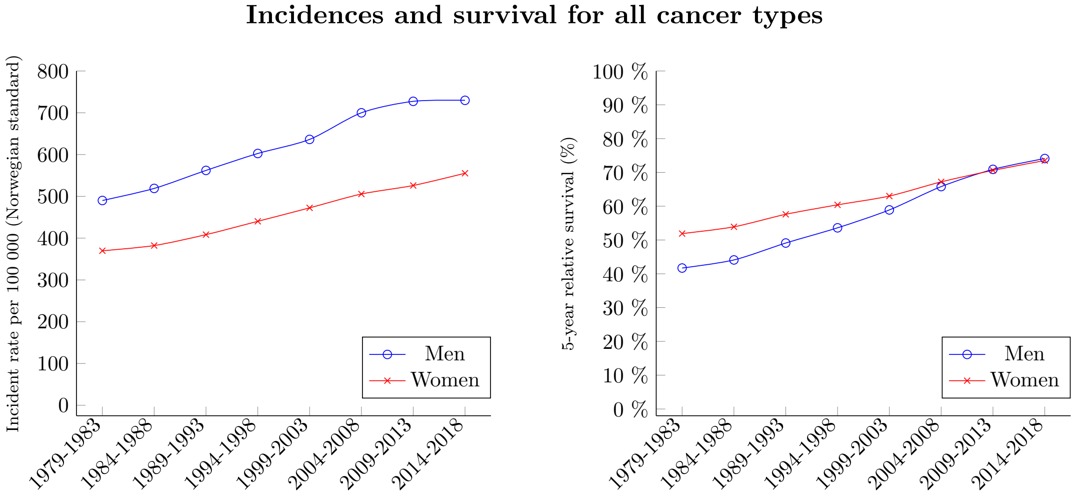


Figure 4: Historical data on cancer incidences and survival in Norway.

Source: Cancer Registry of Norway (2019)

The standardized cancer care pathways (in Norwegian: pakkeforløp for kreft) comprise a fundamental basis for the Norwegian cancer care. The cancer care pathways were introduced in Norway in 2015 and 28 different forms of cancer have an associated CCP (Norwegian Patient Registry, 2020). The CCPs constitute a framework for organizing cancer care and involve all aspects of care, from diagnosis, to treatment to evaluation and control. If the GP suspects the patient of having cancer, the patient should be referred to an associated CCP. The primary purposes of the CCPs are to prevent unfounded waiting times for the patients and quickly provide a diagnosis and treatment. This may be achieved by having a coordinated and integrated approach to health care planning and control

We divide the CCPs into three stages: diagnosis, treatment, and follow-up and control (Figure 5). In the diagnosis stage, the patient is thoroughly examined. If the examination reveals cancer, the patient is referred to the treatment stage. If not, she is dismissed. Figure 6 provides more details of the different phases in the diagnosis stage. In the next phase, the treatment phase, the patient receives her treatment. The form of treatment, e.g. surgery, chemotherapy, radiotherapy or a combination, depends on the outcome of the diagnosis stage. Finally, after the treatment, the patient is provided rehabilitation and monitored closely to reveal any relapses. This thesis concentrates on the two first stages of the cancer care pathway.

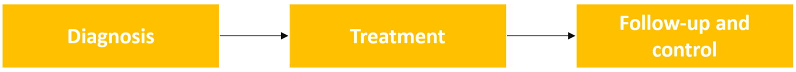


Figure 5: Three stages in a cancer care pathway. In each stage, there are sets of activities that the patient must undertake. These activities depend on the cancer type of the patient.

Source: Norwegian Directorate of Health (2020)

Each CCP provides three indicative time limits in the diagnosis stage, one for each of the three phases. The watches in Figure 6 indicate where in the care process the time limits are found. Phase 1 corresponds to the time from the hospital receives the referral from the GP to the patient meets at the hospital for the first time. The second phase is the time from diagnosis starts till all diagnosis-related activities where the patient must be present are finished and a clinical decision regarding further treatment has been made. The third phase is the time from the diagnosis is established till the treatment starts (Norwegian Directorate of Health, 2020). The time limits are dependent on the phase and the specific cancer diagnosis and range between 4 and 21 days.

The CCPs also describe the following two aspects of cancer care:

- **Medical examination:** The potential cancer illness is identified by examining the patient and her symptoms. If the examination reveals that the patient does not have cancer, she is discharged from the CCP and provided alternative non-cancer related care or no care if she has no illness that needs treatment.
- **Initial care:** The authorization of treatment and preparation. The decision on whether or not the patient should undergo treatment (e.g. surgery, chemotherapy or palliative treatment). The prescribed treatment might require the patient to be prepared, for example by taking a specific medicine before surgery to shrink the tumor (neoadjuvant therapy).

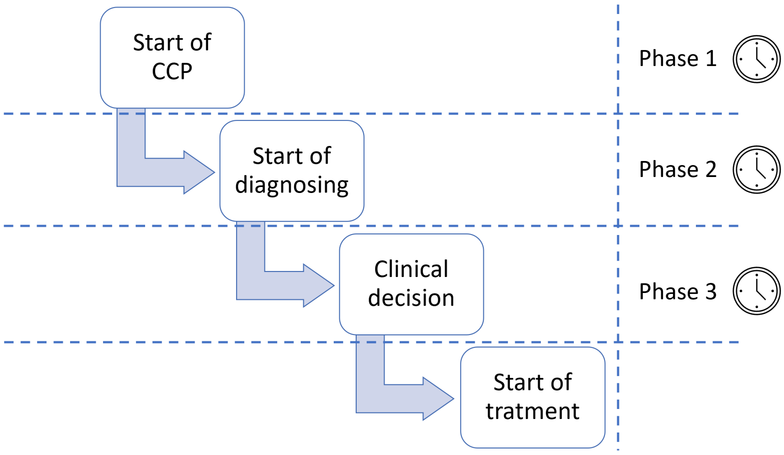


Figure 6: Phases and time limits in the diagnosis stage of a CCP.

Source: Norwegian Directorate of Health (2020)

Two important performance indicators, determined by The Norwegian Directorate of Health, for the CCPs are 1) 70 % of all cancer patients should be enrolled in a CCP and 2) 70 % of the patients enrolled in a CCP should start their treatment within the prescribed time limits. In 2019, the first target was met with 76 % of new cancer patients enrolled in a CCP, according to the Norwegian Patient Registry (2020). However, only a 67 % share of these patients started their treatment within the recommended time limits. The variation in these metrics varies significantly between the different CCPs. For example, 86 % of bladder cancer patients enrolled in a CCP started their treatment on time. The same number for lung cancer patients was 57 % in 2019 (Norwegian Patient Registry, 2020). The three forms of cancer that are assessed in this thesis, namely ovarian, cervical and uterine cancer, are all types of cancer that have low scores on the two performance indicators presented. Data on the two performance indicators for the relevant CCPs of this thesis are given in Table 1.

Table 1: Overview of key performance indicators for gynecological cancer patients in 2019.

	Type of cancer		
	Ovarian	Uterine	Cervical
% of patients enrolled in a CCP	70 %	74 %	70 %
% of patients in a CCP that started their treatment on time	65 %	60 %	56 %

Source: Norwegian Patient Registry (2020)

One of the intentions of introducing standardized CCPs in Norway, in addition to the aspects of predictable and quicker access to cancer care, was to harvest the strengths of multi-disciplinary care. This is among other things done by arranging multi-disciplinary team (MDT) meetings. In MDT meetings related to cancer care, different types of doctors (e.g. radiologists, surgeons, gynecologists, anesthesiologist) and other personnel involved in cancer care (e.g. cancer coordinators and nurses) meet to discuss cancer patients' cases. Often, it is only patients requiring a specially arranged treatment (e.g. due to the physical state of the patient) or borderline cases where there is doubt about which treatment the patient should receive, that are discussed here. The outcomes of these meetings are unified decisions on the treatment for the relevant patients discussed. The reasoning for arranging the MDT meetings is to have multiple approaches to a patient's case, so that all aspects of the patient's health are reviewed in relation to each other. When different specialists meet for discussions, the decisions they arrive upon are not determined solely by one individual, but by the collective of experts.

2.3 Gynecological cancer care at Oslo University Hospital

Preserving and improving people's health is both a global and national concern. But, health care in these perspectives comprises numerous smaller elements. We believe that to start with one of these building bricks, in this case, mid-term planning of cancer care in a single hospital, is a constructive approach to solving a very complex and large problem. In this section, we give a brief overview of cancer care at Oslo University Hospital today at the Department of Gynecological cancer.

Gynecological cancer is defined as malignant tumors originated in the female genitals. Different types include cancer in the vulva, vagina, uterine, ovarian, tuba uterina and cervix. 1747 women were diagnosed with gynecological cancer in Norway in 2018 (Kristensen et al., 2017). The most prevalent types of gynecological cancers are uterine cancer, ovarian cancer and cervical cancer. When referring to gynecological cancer in the rest of this thesis, it is referred to cancer in one of these three areas. The majority of the patients treated at the Department of Gynecological cancer at Oslo University Hospital has cancer in one or more of these sites. According to the left plot in Figure 7 the incidence rate of gynecological cancer has remained stable over the last 40 years, with approximately 60 incidences per 100 000 Norwegian women each year. Overall, the incidence rate of cancer for women is 550 per 100 000 women, meaning that approximately one in ten cancer incidences among Norwegian women occurs in the genitals (Cancer Registry of Norway, 2020). The survival rates for gynecological cancers have increased slowly with recent numbers showing that over 70 % of all gynecological cancer patients are alive five year after they are diagnosed. Norway has the highest worldwide survival rates for ovarian and cervical cancer (Allemani et al., 2018).

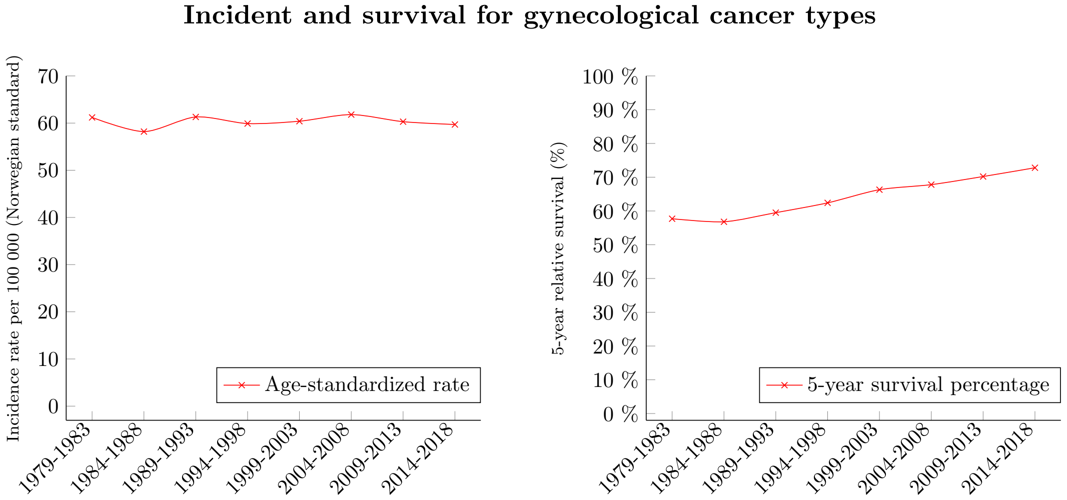


Figure 7: Historical data on cancer incidenses and survival for ovarian, uterine and cervical cancer in Norway.

Source: Cancer Registry of Norway (2020)

The CCPs for gynecological cancer types usually start at the GP's office where the patient is referred to the hospital for further investigation if cancer is suspected. From the admission office, the patient is given an initial appointment at the gynecological cancer outpatient clinic. Here, the patient speaks to a nurse or gynecologists that tell her about the further steps in the CCP and which activities she must undertake. The hospital may also have received test material from the GP, for example a biopsy or blood sample, that is analyzed at the hospital's medical laboratory. Then, the diagnosis phase starts. Typical activities that must be undertaken are gynecological examination, CT scan, MRI scan, rectoscopy, cystoscopy and positron emission tomography (PET) scan. After the diagnostic tests are performed, the patient is diagnosed, and the treatment may start. If the patient has cancer, a conversation with the patient is scheduled to provide information to the patient about the proposed treatment. This conversation is carried out by a nurse at the cancer clinic and may be performed using the telephone. The treatment may only start after the patient has received information about the proposed treatment. The treatment, and implicitly the type of resources and personnel needed to treat a patient, is known with certainty only after the diagnosis for the relevant patient is set. The type of treatment a patient receives is dependent on the type and stage of the cancer. The main types of treatment for the gynecological cancer types are radiotherapy, surgery and chemotherapy.

It is noted that time limits in the CCPs are only defined for the diagnosis stage, that is up until start of treatment, but that the CCPs cover all aspects of cancer care, also after the treatment is finished (Norwegian Directorate of Health, 2020). An example on what the CCP up until start of treatment for a gynecological cancer patient may look like, is provided in Figure 8.

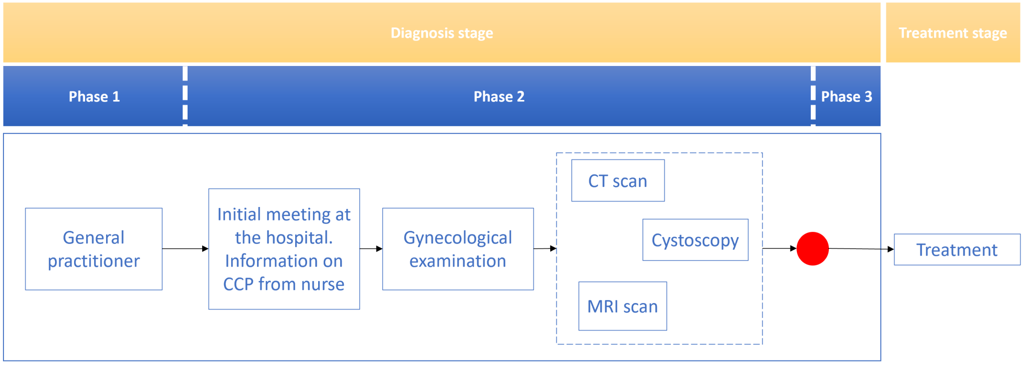


Figure 8: Example of a CCP for a gynecological cancer patient up until start of treatment. The red dot indicates the decision on which diagnosis the patient has and her treatment. The dashed line around the activities CT scan, cystoscopy and MRI scan indicates that the order of execution for these activities is unimportant.

The operational planning of cancer care is carried out by cancer coordinators. The cancer coordinators are specialized personnel, often nurses, that are employed by the hospital to ensure that each patient is given the right care at the right time and prevent unnecessary waiting time (Oslo University Hospital, 2019). The coordinators book appointments for the patients, keep the individual patients informed and coordinate activities between hospital units. Coordinating across hospital units at Oslo University Hospital is sometimes demanding due to the different planning practices and systems in different units. Some units prefer appointments to be booked manually in spreadsheets in pre-defined time slots, while other prefer other forms of manual communication like phone, email or even fax. They are also responsible for finding solutions for the patient that are not able to follow their CCP due to, for example, the physical state of the patient or comorbidities. The flexibility that the cancer coordinators provide for the patients ensures patient safety on a patient level. At the same time, this flexibility comes at a cost, and increased flexibility for the individual patient might go at the expense of the timeliness of care for the collective patient population.

The tactical planning at the Department of Gynecological Cancer is carried out by the unit's management and planners in cooperation with the attending physicians and planners from interrelated hospital units. The process is repeated with a certain interval, or if there are abrupt shifts in demand that require an updated allocation of resources. This planning process, which in practice is a process of resource allocation between patient groups, is based on historical data of admitted patients per patient group, negotiations and reward schemes. The historical data forms the blueprint for the mid-term resource allocation policy. That said, last year's demand for one activity is not enough to predict this year's demand. For example the prevalence of one type of cancer type might be higher one year compared to earlier, or queues might have built up due to unforeseen events. Negotiations and prioritizing must therefore take place to adjust the blueprint. Last, but not least, the reward schemes, which were mentioned in Section 2.1 also apply directions to the planning process by motivating certain behavior. The reward schemes are in many cases not designed by the hospital itself, but by politicians and lawmakers, making the process of allocating resources a centralized political decision. This is in opposition to the message we proclaim, namely to take on an integrated approach to multi-disciplinary planning and control of health care.

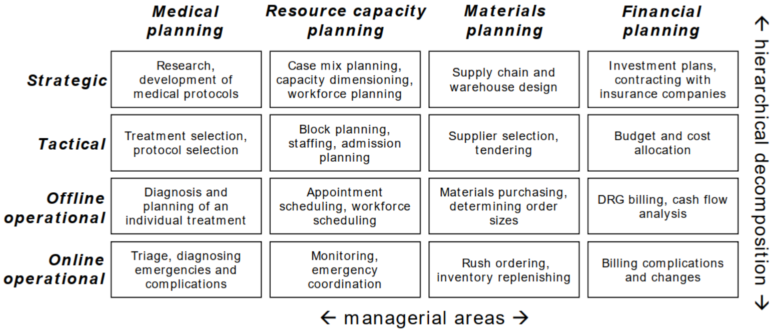
3 Literature review

This chapter provides a concise review of the relevant literature for the scientific setting. Articles considered are mainly on prescriptive operational research methods, such as simulation and mathematical programming in health care. The search is narrowed down to articles on tactical multi-disciplinary planning, with the main focus being on resource capacity planning and patient admission planning. Both deterministic and stochastic approaches to variability are discussed.

Literature on multi-disciplinary planning in health care is strongly related to existing literature on appointment scheduling, where patients are allocated capacity at a single resource (Marynissen and Demeulemeester, 2019). In multi-disciplinary planning, problems involve multiple patient groups that are to be assigned multiple resources over multiple time periods. The first contribution to this field of study can be traced back to the work of Bailey (1952), who modeled an outpatient appointment and queuing system. In the following years and decades, many contributions have been made to this field of study. We refer to Hulshof et al. (2012) and their taxonomic classification of articles on health care in operations research and management sciences. For more comprehensive reviews of literature on multi-disciplinary planning in health care in particular, we refer to the contributions of Vanberkel et al. (2009) and Leeftink et al. (2018).

This thesis considers planning on a level with a medium-long planning period. Hans et al. (2012) propose a framework for planning and control in health care. The framework works as a tool to structure and systematize health care functions. Figure 9 depicts the concluding matrix of the framework. According to the framework, strategic planning refers to long-term planning where decisions may embrace longstanding research decisions, procurement contracts and supply chain design. Tactical planning refers to mid-term planning, meaning that decisions usually are applicable for the coming weeks or months. Planning and control with a planning period of a few days or shorter is denoted operational planning. Online operational planning deals with monitoring a process and acting instantaneously on unforeseen events. Examples of online operational scheduling are rush-orders of hospital materials and triaging. Offline operational planning refers to short-term planning carried out in advance of the execution of an operation, examples are short-term nurse rostering and treatment selection.

With respect to the framework for health care planning and control presented by Hans et al. (2012), the planning problem of this report, which is described in Chapter 4, is placed within the *tactical hierarchical level*. The relevant managerial area for this thesis is *Resource capacity planning*. This managerial area considers dimensioning, planning, scheduling and control of renewable resources, e.g. operating theaters and MRI scanners.



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

Figure 9: Framework for health care planning and control.

Source: Hans et al. (2012)

The rest of this chapter is divided into five sections. In each section one subject is reviewed in relation to this thesis and relevant articles are brought up for discussion. Section 3.1 describes related objective functions, Section 3.2 describes suitable solution methods, Section 3.3 presents literature on approaches to different types of systems in terms of job shop scheduling and Section 3.4 discusses different approaches to variability. The chapter is summarized and positioned relatively to existing literature in Section 3.5. An overview of the most important papers mentioned in this chapter are compiled in Appendix A.

3.1 Objective function

The objective function of the mathematical model should reflect the most important performance indicator(s) of the system. When several resources and multiple patient groups are modeled, a method to meet the demands and expectations of several different stakeholders is often preferred. The objective function in an optimization problem for a multi-disciplinary planning problem in health, could for example reflect the need for patients to minimize waiting times, the hospital's ambition to maximize resource utilization or a combination of the two. Different approaches to the objective function can be made, but we distinguish between optimization with multiple objectives (multi-objective optimization) and optimization with one single objective (single-objective optimization).

A challenge with using multiple objective functions in a planning problem, is to decide how to rank solutions. Determining which solution is the best, might be as challenging as finding satisfying solutions. On the other hand, single objective functions struggle to see the broad picture. Here, one performance indicator is assessed in isolation and its influence on other indicators is downplayed. Deciding what the objective function should look like must not be taken lightly, as it affects the output of the model, its complexity and consequently, its solution method. In this section, we briefly mention which solution method that is being used in each article. For a more thorough assessment of relevant solution methods, we refer to Section 3.2.

In this section, we assess literature on both single- and multi-objective optimization, due to the widespread use of both in the context of multi-disciplinary planning. Single-objective and multi-objective optimization are discussed in Sections 3.1.1 and 3.1.2, respectively.

3.1.1 Single-objective optimization

The objective function motivates the system to perform in a certain manner, and failing to propose an objective that reflects the purpose of the system, may be the difference between success and failure. Choosing the ”right” objective is not done based on some standardized rule, but depends on the problem that is to be solved. Azadeh et al. (2014) study the scheduling of patients in an emergency department laboratory. They propose a single objective function which aims at minimizing the total weighted completion time of all patients. The values of the weights are set in a manner that provides sufficient care to urgent patients. But, the model does not consider the next steps of the patient care process, such as treatment and control. They suggest to use a genetic algorithm and a response surface methodology to solve the scheduling problem. Hulshof et al. (2013) proposed a similar objective function when modeling an integrated model for care chain planning. In contrast to Azadeh et al. (2014), they suggest a weighted objective function that minimizes the number of patients waiting in queues, and not the time these patients have spent in the queue. A similar objective functions is suggested by Castro and Petrovic (2012), where the number of patients exceeding waiting time targets are minimized. Chern et al. (2008) propose a two-phased binary integer programming model with the objective to minimize waiting times for both patients and doctors.

3.1.2 Multi-objective optimization

Petrovic et al. (2013) address a tactical planning problem in radiotherapy appointment scheduling. They introduce an objective function that combines the minimization of average waiting times for all patients, percentage of patients that do not meet the required due date for the first treatment and employee overtime used. This combination of objective functions is able to optimize for three different metrics that are crucial in order to obtain a functioning multi-disciplinary system. To obtain an optimal, or close-to-optimal, solution, they propose a genetic algorithm. They find that genetic algorithms are well-suited to explore the vast solution space of appointment scheduling problems in a health care setting and are able to find acceptable solutions to real life-sized problem instances in reasonable time. A weakness of this genetic algorithm is that it requires manually inputted objective function weights to determine which solution is the best. This means, that defining what a "good" solution is, is of subjective nature. A more intelligent approach, would be to use the ϵ-constrained method. Using this method, each objective is transformed into a single objective problem (Section 3.1.1), modifying the other objective functions to become constraints that are constrained by a value ϵ. Solving this problem for each of the objectives iteratively, may yield the Pareto optimal front. An application in a health care setting, although not multi-disciplinary, of this approach is found in Gullhav et al. (2018).

Jerić and Figuiera (2012) study appointment scheduling of medical treatments for resident patients in a hospital. They propose a multi-objective objective function where the different objectives are assigned weights. The objectives are to maximize the number of treatments, minimize the maximum total possible waiting time for physicians, minimize physician busyness and minimize the number of time periods where critical equipment is occupied. The weights of the objective functions are decided upon by the mathematical model. To search for the optimal combination of weights for the objective function, they test the performances of variable neighborhood search algorithms, scatter search algorithms and non-dominated sorting genetic algorithms. Their work concludes that a scatter search algorithm with randomized combination of solution is the most suitable heuristic approach to reach a satisfying solution. This multi-objective formulation is composed by four objective functions. At one hand, a formulation on this form enables the solution to capture different aspects of the problem that would not be captured by a single objective. On the other hand, multiple objectives increase the complexity of the model, and implicitly, increase the runtime. As Jerić and Figuiera (2012) underline, the problem is practically unsolvable if heuristics are not applied in the solution process.

3.2 Solution method

Relevant literature on multi-disciplinary planning often encompasses both an analytical mathematical model and a discrete event simulation model. In this section, we briefly introduce and discuss different solution approaches that are made in the context of multi-disciplinary planning in a health care setting. In Section 3.2.1, exact solutions using MIP solvers are discussed and Section 3.2.2 encompasses the use of simulations. Finally, in Section 3.2.3, other solution methods are discussed.

3.2.1 Exact solution using MIP solver

With respect to an exact solution by a MIP solver, Gartner and Kolisch (2014) present a model for patient flow in a German hospital. They use a rolling horizon approach to solve the problem, and the objective is to maximize the contribution margin. It is shown that relatively large test instances can be solved to optimality within a fraction of a second on average. Hulshof et al. (2013) propose an integrated method to decision-making for multiple resources, time periods and patient groups. The model, which is modeled using a rolling horizon approach, is almost solved to optimality within minutes for relatively large test instances. Both papers have a low level of detail in the modeling, and do only consider flow of patients and not individual patients, nor any form of uncertainty.

In the context of scheduling elective patients, Conforti et al. (2011) suggest a model that is solved to optimality using a MIP solver. The model makes decisions on which patients to admit and when to admit them. The objective is to maximize the weighted number of patients admitted to the health care system. The authors include beds for inpatients and block planning in the scheduling problem. They propose a model where a patient's length of stay maximum can be five consecutive weekdays. This imposes heavy restrictions on time limits and capacities, but forces the length of stay to a minimum. A similar approach is made by Saadani et al. (2014) who models a resource allocation problem for multiple resources, solved to optimality using a MIP solver for smaller instances, with the objective to minimize stay-durations for patients.

3.2.2 Simulation models

In order to evaluate a solution, generate new solutions, handle uncertainties or assess different allocation policies, simulation is a powerful tool. Using simulation to verify the feasibility of a solution in a health care setting is popular among researchers due to the ease one can test a complex model with a large state and decision space in reasonable time with (Leeftink et al., 2018). A widely used generic framework for formulating simulation models is presented by Hillier and Lieberman (2015). They suggest that a simulation model of a system should consist of the following building blocks: a definition of the state of the system, an identification of the possible states of the system, an identification of possible transitions from one state to another, an identification of possible events, a simulation clock and a method for randomly generating events.

Simulations provide a way of testing a model's appearance in a real-life setting by only using a computer. Of course, it has its flaws. For example, when generating random events, assumptions regarding probability distributions must be made. Assuming that a certain process follows a specific probability distribution is a strong assumption, and it is noted by Hillier and Lieberman (2015) that the results of a simulation only provide estimates of the performance of the system, and caution should be exercised before drawing any rigorous conclusions.

The majority of the articles that apply simulation in a multi-disciplinary health care setting use it for evaluation purposes. A common approach is to first solve the problem using a set of pre-specified configurations, and then test these solutions under uncertainty (Bikker et al., 2015; Cardoen and Demeulemeester, 2009; Liang et al., 2015; Romero et al., 2012). The goal of this type of simulation is to assess how a solution of an optimization model, often solved with deterministic input values, reacts to different types of uncertainty, and evaluate how well the solution performs in a more realistic setting. The most prominent types of uncertainties regarded in the health care setting are presented in Section 3.4.

Pérez et al. (2013) propose different allocation policies at a multi-step nuclear medicine department. They do not propose any objective function explicitly. However, different objectives are implicitly built into the resource allocation policies in terms of patient prioritization rules. Dharmadhikari and Zhang (2013) use simulation-based optimization to evaluate the effectiveness of different block scheduling policies in a multi-clinic system. They found that a first in, first out (FIFO) scheduling policy was sufficient in scenarios with low demand, but when demand increased, the attributes of the patient should be taken into account when scheduling patients.

Addis et al. (2015) simulate an operating room scheduling and re-scheduling problem using a rolling horizon approach. At the beginning of each week, a period of several weeks is scheduled, but only the first week of the generated schedule is applied. They keep the remaining schedules as a reference for following iterations. They also incorporate robust optimization in order to handle uncertainty linked to surgery times that are hard to predict.

It is noted that classifying simulation as a solution method may sometimes be misleading. As proposed above, simulations are often utilized when evaluating the performance of a solution. The simulation model itself does not propose a solution to the problem, nor does it make any decisions. However, simulations do operate in close cooperation with optimization methods. Different approaches to integrate the two have been made, and a taxonomy on hybrid simulation-optimizations is proposed by Figueira and Almada-Lobo (2014). Hence, we have positioned material on simulation in the solution method section.

3.2.3 Other solution methods

Heuristics

The motivational factors for solving multi-disciplinary planning problems using heuristics are stated by Leeftink et al. (2018). They state that one of the most important reasons for utilizing heuristics as solution method, is to handle the complexity of the planning problems. Tactical scheduling problems are in many cases NP-complete (Afroze and Gardell, 2015), meaning that finding the optimum is difficult. Heuristics are often able to provide a satisfying, though not guaranteed optimal, solution, within a shorter amount of time than approximation algorithms, decompositions and reformulations.

Du et al. (2013) propose a hybrid genetic algorithm for solving a scheduling problem of clinical patient pathways. The algorithm combines particle swarm optimization (Kennedy and Eberhart, 1995) and a genetic algorithm. The strength of this approach is that population diversity is maintained and premature convergence prevented. Du et al. (2013) suggest that the process of tuning parameters could be eased through implementation of local search heuristics. Petrovic et al. (2013), Azadeh et al. (2014) and Azadeh et al. (2015) use genetic algorithms to solve their multi-disciplinary planning problems. Saremi et al. (2015) use a multi-agent tabu search algorithm to solve a bi-criteria optimization model of a multi-stage facility.

Jerić et al. (2011) study the scheduling of resident patients in a hospital using binary integer programming. They introduce a reduced variable neighborhood search (RVNS) heuristic. The regular variable neighborhood search approach looks for the optimum among points in a predetermined number of neighborhoods. The RVNS on the other hand randomly changes the neighborhood throughout the search, which prevents it from being trapped in a local extremum. This enables larger problem instances to be solved by exploring more of the solution space in an efficient manner (Hansen et al., 2019).

Other solution methods

Barz and Rajaram (2015) address an elective patient admission and scheduling problem with multiple resources and patient groups where the objective is to maximize net contribution. The problem is modeled as a Markov decision process, but simplified using approximate dynamic programming techniques due to issues with complexity. The new simplified problem yields upper bounds to the problem, and these are utilized in a new optimization model in order to find the marginal values of the constrained resources in the original problem. Finally, these marginal values decide the prioritization rules when scheduling patients. Leeftink et al. (2019) study multi-disciplinary outpatient clinic planning with stochastic patient care pathways. Their objective is to design blueprint schedules for clinicians. To handle stochasticity, they utilize a sample average approximation algorithm. Using this method with a sufficiently large number $|N|$ of randomly drawn scenarios, a robust solution is acquired. Hulshof et al. (2015) use stochastic programming and ADP to allocate resources in a multi-disciplinary setting to minimize the patients waiting for a treatment in a hospital.

Other solution methods that are applied in the context of multi-disciplinary resource planning in hospital settings are constraint programming (Hahn-Goldberg et al., 2014), Lagrangian relaxation (Sadki et al., 2011), hierarchical programming (Castro and Petrovic, 2012).

3.3 Type of system

In this section, different types of systems with regard to precedence constraints are introduced, and articles using the different approaches are referred to. Precedence constraints are constraints that impose restrictions on the order of activities a patient must undertake, for example: a CT scan and MRI scan with interpreted images must be present before an MDT meeting can take place. To clarify which type of system one deals with in a multi-disciplinary planning problem, is decisive for how to model and in turn solve the problem. The three types of systems presented in this section, and visualized in Figure 10, represent different optimization problems (Leeftink et al., 2018). We distinguish between three different system types: flow shop, open shop and mixed shop. The three types of systems are presented in Sections 3.3.1, 3.3.2 and 3.3.3, respectively.

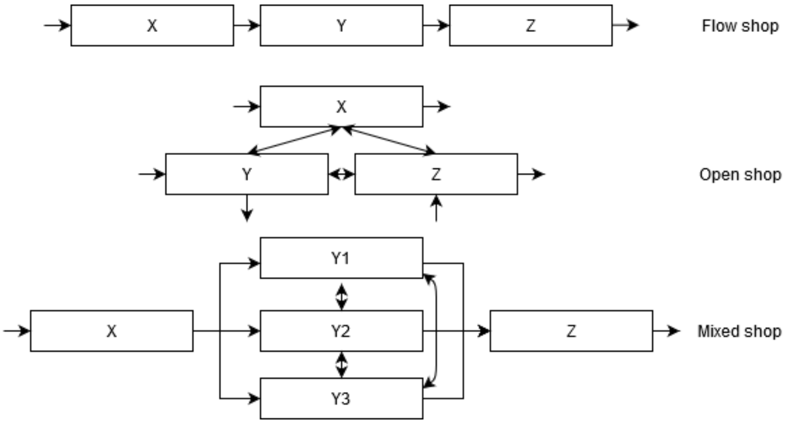


Figure 10: Three different types of system: flow shop, open shop and mixed shop systems.

3.3.1 Flow shop

A flow shop system is a system where the patients' sequence of activities are pre-defined, meaning that there are strict precedence constraints. This allows for a high degree of standardization of services, but limits flexibility. This type of system is typically put to use in specialty clinics, that is, clinics focusing on a defined group of patients where there are only small variations in the procedures for each patient. Alfonso et al. (2011) utilize flow shop scheduling in their paper where a blood collection site is modeled. Using flow shop precedence constraints, von de Vrugt et al. (2017) model the diagnosis services at a Dutch breast cancer center.

3.3.2 Open shop

In an open shop type of system, each patient undergoes activities in an order which is not fixed. The flexibility is high in this system. Process divergence, that is the possible pathways the patient may take, becomes large since there are no, or at least very few, restrictions on precedence between activities. A cancer patient that must undergo a CT scan, MRI scan and a gynecological examination in an arbitrary order as part of the diagnosis phase, may be modeled using an open shop system. Examples on open shop type of systems are Azadeh et al. (2014) who model patients in an emergency department laboratory, Vermeulen et al. (2007) that address a generic multi-appointment and multi-resource scheduling problem and Matta and Patterson (2007) who model an oncology center as an open shop system, and evaluate its performance when the scheduling practices are changed.

3.3.3 Mixed shop

According to Leeftink et al. (2018), a mixed shop system includes precedence constraints for activities, but with flexibility when sequencing some of the activities. A practical example of a mixed shop type of system is if a patient first requires an intake consultation, then a set of activities, which all have to be performed, but without precedence constraints (e.g. biopsy, CT scan and MRI scan), and then a patient-specific treatment with precedence constraints. If we were to refer to Figure 10, *X* is equivalent to the intake consultation, *Y1, Y2* and *Y3* corresponds to the activities where there are no precedence constraints and *Z* is equal to the patient's treatment stage.

Examples of mixed shop systems are found in Conforti et al. (2011) who address a patient-centered week hospital planning problem, Chern et al. (2008) that present a scheduling algorithm for health examination packages in a hospital and Saremi et al. (2015) who introduce an appointment scheduling problem with heterogeneous service sequences in a multi-disciplinary facility, that is, all patients act as mixed shop-agents that require service in a deterministic, heterogeneous sequence. The work of Saremi et al. (2015) on patients with heterogeneous service sequences are based an article by Pham and Klinkert (2008), but the latter does only take into account surgical case scheduling.

3.4 Variability approach

In order to model a physical problem accurately, variability, or uncertainty, should be taken into account. The world as we know it is not deterministic, and optimization models should reflect this when necessary. This is particularly important when modeling a multi-disciplinary system where uncertainty in one part of the system propagates quickly to other parts of the care chain due to interrelatedness between hospital resources and units (Leeftink et al., 2018). This propagation of uncertainty is closely related to the bullwhip effect introduced by Forrester (1961). This effect is thoroughly assessed in the literature on industrial applications (Badar et al., 2013; Zotteri, 2013; Wang and Disney, 2016).

The characteristics of the uncertainty in a problem are related to the hierarchical planning level of the problem (Hans et al., 2012). To plan for uncertainty in a long-term planning problem is fundamentally different from planning for uncertainty on an operational level. Long-term strategic decisions may be exposed to more uncertainty since the implementation of the decision happens long after the decision. Also, the type of system (Section 3.3) is relevant when assessing uncertainty. For example, a mixed shop type of system is better suited to handle uncertainty since it has flexibility that the flow shop type of system misses. Below, we briefly introduce four highly relevant aspects of a multi-disciplinary system where uncertainty might be present:

- **Patient arrivals:** The uncertainties in patient arrivals mainly arise due to patient no-shows and late arrivals. Predicting the rates of these may be approximated using probability distributions and historical averages. Among the probability distributions, the Poisson distribution is historically the most popular as arrival processes often tend share its shape. It is also a discrete probability distribution, which makes it useful in real-life applications when we deal with an integer number of patients (Green, 2006). Papers that take into account stochasticity in patient arrivals in multi-disciplinary planning problems are Romero et al. (2012), Bikker et al. (2015) and Leeftink et al. (2016).
- **Appointment durations:** Due to large differences in required care from patient to patient, even for the same type of appointment, appointment durations are subject to large uncertainties. To handle this type of uncertainty is especially important in health care systems where patients are scheduled for multiple appointments in a single day. Kalton et al. (1997) and Liang et al. (2015) model stochastic appointment durations.
- **Resource capacity:** Uncertainty in resource capacities arise due to seasonal effects (e.g. higher demand in some periods of the year) or due to random events, such as machine breakdowns or key employees calling in sick, according to Leeftink et al. (2018). The best way to control this type of uncertainty, is to apply forecasting techniques based on historical data. It is noted that planning for this type of uncertainty using stochastic programming is very hard and few papers have attempted it.
- **Care pathway:** If a patient's treatment is heavily impacted by results from the diagnosis phase, the model should include the uncertainty of the care pathway. This is important in many types of health care - for instance in cancer care, where the results from each step of the care pathway influence the subsequent steps. Examples of papers modeling uncertainty in the care pathways are found in Cardoen and Demeulemeester (2009) and in Venkitasubramanian et al. (2015)

Not taking uncertainty into account influences the robustness of the model. On the other hand, introducing uncertainty will in many cases lead to an exploding state space and long runtimes. It is therefore necessary to determine which uncertainties that are present in the problem and which of these that are of an importance that qualify them to be included in the optimization model. To simplify calculations, authors often optimize problems deterministically, but evaluate and test them in a stochastic environment using simulations (Leeftink et al., 2018).

3.5 Positioning of this thesis

In this section, a brief summary of existing gaps in the relevant literature on multi-disciplinary planning is given. Finally, this thesis' contribution to fill in these voids is specified. Important voids in the literature are:

1. Current approaches to planning in health care are able to provide optimal or near optimal solutions for highly detailed problems that encompass one resource and multiple patient groups, or multiple resources and one patient group. However, there is a lack of methods to solve planning problems that includes multiple resources and multiple patient groups in an effective manner.
2. Timely care is important in many sorts of care, especially cancer care. In existing literature time limits are rarely devoted attention, often due to issues related to complexity and heterogeneous service sequences.
3. Many studies consider optimization problems with a strictly limited planning period, often only a couple of days to a few weeks. What these studies lack, is the ability to look ahead in time and evaluate the model's performance and dynamics over a longer period of time.

To undertake the gaps in the existing literature, we propose an optimization model with a weighted objective function that aims at minimizing the waiting time between activities for patients in a care setting at a hospital. The model is to be solved using a MIP solver and thoroughly tested and evaluated using simulations.

This thesis complements to already existing work in the following ways: first, in many types of care, and cancer care in specific, quick access to care is vital. In our work, this is ensured by introducing strict restrictions on when selected key activities in a care process must start within. The model proposed is flexible in the sense that not all activities need to have an associated time limit.

Next, our model is tested and evaluated using a rolling horizon approach in the simulation to observe the dynamics of the system over multiple planning periods. By simulating a longer time span, we are able to provide a more accurate and realistic replication of the system that is modeled. The rolling horizon approach also contributes to ensure predictability in the planning process, by restricting the number of changes in the schedule for overlapping time periods from planning period to planning period. The rolling horizon approach is elaborated on in the Chapter 5.

Third, our work offers an objective function, inspired by Hulshof et al. (2013), that is interpreted as the weighted amount of days patients have been waiting in a queue. The weights in the objective function may be set manually or based on a specific scheduling policy to motivate the system to behave in a certain manner. In contrast to the objective function of Hulshof et al. (2013), our objective function enables the hospital to assign weights based on how long a patient has waited in a specific queue, but also to track the total number of time periods the patient has been waiting. The model is designed so that the objective function may be replaced effortlessly to optimize other relevant performance criteria, such as minimizing the makespan or maximizing the number of admitted patients.

In summary, our contribution is in contrast to existing literature able to assess the problem of allocating multiple resources to multiple patient groups, or care processes. We are also able to evaluate the problem over multiple planning periods in order to capture the system dynamics over a longer time span than, using optimization in close cooperation with simulation.

4 Problem description

In this chapter we describe the problem of patient admission and resource planning for patients enrolled in a standardized care process in a multi-disciplinary care system. By keeping track of queues linked to each hospital activity, resources are allocated to the correct activity at the right time, ensuring sufficient and timely care. When the problem is solved, schedules for patient admission and resource allocation are generated. Decisions must adhere to pivotal restrictions regarding time, capacity, sequencing and queues.

We distinguish between a planning period and an implementation period. The planning period is the set of time periods we generate plans for when running the optimization model, whereas the implementation period consists of the time periods where a realization of the system takes place and parts of the schedule generated in the planning period are implemented. The optimization model is run with deterministic mean values, but the realization of the schedule is subject to uncertainty. In this thesis, the implementation at the hospital is replaced by a scheduling framework with a simulation cooperating closely with the optimization model. The scheduling framework is presented in Chapter 7.

First, in Section 4.1 a motivating example that illustrates key aspects and dynamics of the problem is presented. The example is inspired by a patient enrolled in a cancer care process, which is a typical standardized care process. Although the example is on cancer care in specific and that we have collaborated with a hospital department specialized in cancer care in the preparation of this problem, we have formulated the problem description generically. This is to ensure that the model also could be utilized in a non-cancer context. Next, the problem description is provided in Section 4.2. Here we describe the purpose of the problem, available information, relevant restrictions and fundamental decisions that are to be made.

In Section 4.3, we assess the dynamic and uncertain context of the problem. It is of high interest to assess how the system behaves over multiple planning periods. The planning problem itself, as described in this chapter, remains unchanged from planning period to planning period, but its starting condition, for example the number of patients that reside to the system, changes each time the optimization model is run. We also assess uncertainties that arise when the optimization model is realized in a real-life environment.

4.1 Motivating example

A short example is provided in the following to better illustrate the problem. Mia, a 57 year old woman, is not feeling well and suspects that something might be wrong. She calls her GP and books an appointment in two days. The GP takes some blood and urine samples and examines Mia. The visit at the GP marks the beginning of Mia's care process. After a few days, the GP calls Mia back to tell that he suspects Mia to have uterine cancer. The GP has already notified the hospital. The hospital contacts Mia and grants her an appointment for further examinations at the hospital in three days. Mia is now enrolled in a cancer care pathway associated with uterine cancer.

At the hospital, Mia is given information about possible outcomes of the diagnosis stage and that the diagnosis stage for uterine cancer consists of three activities that must be undertaken: CT scan, biopsy and a gynecological examination. Each of the activities requires one or more resources to be allocated to them in order to be carried out. For example, the CT scan activity needs a CT scanner, a radiologist and two radiographers for preparation and interpretation, the biopsy requires an amount of pathologist time and lab resources and the gynecological examination requires an examination room, a gynecologist and a nurse to be available. All activities in the diagnosis stage must be undertaken before a decision on further treatment can be made.

Mia is told that she is assigned a slot in both the CT scanner today, but that she has to queue for a gynecological examination and biopsy. She is told that her diagnosis will be ready in maximum 12 days. For each activity in a care process, there is a queue.

After the gynecological examination, the diagnosis phase in the uterine cancer care pathway is over. The results from the diagnostics indicate that Mia has a minor tumor in the uterine that could be removed by surgery. She is immediately notified and called to the hospital the next day to plan further steps in the care process. At the hospital, Mia is informed that the physicians recommend a surgery followed by chemotherapy. Based on Mia's diagnosis, uterine cancer, a date for her surgery is set 13 days ahead in time. Simultaneously, she is queued for a new gynecological examination after the surgery to assess the result of the surgery and to plan the chemotherapy. This examination may at earliest be scheduled seven days after the surgery due to Mia's physical health.

Mia undergoes the surgery and the following gynecological examination. Shortly after the gynecological examination, the chemotherapy starts. After finishing the chemotherapy, Mia is enrolled in the follow-up and evaluation program where she periodically revisits the hospital to take tests to ensure that the cancer is under control.

4.2 Problem description

In this section, we describe the problem for one planning period only in a deterministic setting. The behavior of the problem over multiple planning periods and under uncertainty, is discussed in Section 4.3.

4.2.1 The objective of the problem

The main goal of assessing the problem of tactical patient admission and resource planning for patients enrolled in a multi-disciplinary care system, is to provide care for the relevant patients and ensure that waiting times for the patients is held at a sufficiently low level. To motivate the fulfillment of the goal, a suitable objective to the problem is to minimize the total waiting times for patients waiting in a queue for an activity in their respective care processes.

4.2.2 Information available

Patients

The patients included in the planning problem are referred to the hospital by their GP with suspicion of an illness that requires care from multiple hospital disciplines (e.g. cancer). We assume that the expected number of patients referred to the hospital per time period with the suspicion a cancer diagnosis is known. Each patient must follow one care process. Concerning the planning of cancer care, we assume that a care process is equivalent to a CCP. The complete set of possible care pathways is assumed to be finite and known.

Each care process is divided into a diagnosis stage and treatment stage. Which diagnosis process each patient follows in its diagnosis stage is assumed known upon arrival and depends on which illness the GP suspects when referring the patient to the hospital. The treatment stage is dependent on the results from the diagnosis stage. It is assumed that the transition from the diagnosis stage to the treatment stage acts in accordance with a probability distribution based on empirical data of which treatment that follows after a specific diagnosis stage. The transition from the diagnosis stage and the treatment stage is the only point in the care process where a split may happen. That is, we assume that within the diagnosis stage and within the treatment stage, the order of activities is known. The patients may leave the system if no illness is revealed after the diagnosis stage.

Most patients enter the hospital after being referred by their GP and have their entire care process carried out at the same hospital. But, the problem also includes patients that only require some activities to be carried out at the hospital we plan for, e.g. they have their diagnosis ready from another hospital and only require the treatment stage at our hospital. These patients are referred to as external patients, since they arrive from sources external to the hospital in focus.

Activities

The problem consists of a set of hospital activities, e.g. CT scan, MRI scan and gynecological examination. Each care process consists of a subset of the hospital activities that must be undertaken. It is assumed that patients show up at the hospital on time when they have a scheduled activity. The expected length of each activity, i.e. the service time, is known and assumed to be the same for all patients undergoing the same activity.

Queues

We assume that patients must wait in a queue before being serviced at an activity. We let each queue be defined by an activity and a care process, meaning that the different queues a patient must wait in are defined by her care process and its associated activities. This implies that patients residing in the same queue are homogeneous. We assume that the queues do not start empty when the planning starts, and that there are patients inhabiting the queues before the planning is done, i.e. before the first time period in the planning period.

Multiple activities for a patient may be scheduled in the same time period, unless there are special restrictions forcing a patient to wait a pre-defined amount of time periods before being allowed to enter the next activity in the care process. If no such special restrictions apply, the patient is moved to an open slot in the queue for her next activity immediately after begin serviced at the previous activity. Patients may stay in a queue a finite number of time periods before they must be serviced. The hospital chooses a scheduling policy, for example first in, first out, where a patient's position in a queue is solely determined by when she entered the queue, or choosing the patient that have the highest total waiting time in the system. It is noted that patients do not have to be physically present when waiting for an activity in a queue. In Section 5.4, the dynamics of the queues are visualized and elaborated on.

Resources

Let there be a set of hospital resources associated with the execution of the activities in the care processes. To perform an activity, an amount of one or more resources must be made available (e.g. a radiologist and an MRI scanner must be available to perform one MRI scan and the following interpretation). It is assumed that each patient in the same queue demands the same expected amount of resources for the same activity. Since a queue is defined by a care process and an activity, there exists a collection of queues associated with the same activity, and in turn, the same resources. Some resources, for example the CT and MRI scanners, are shared between all hospital departments. Other resources are specialized at serving one type of patients, for example an operation suite only for use by gynecological cancer patients.

The resources have a set of time periods where their capacities may be reserved. A resource's availability is given as a finite amount of time units per time period. The total amount of resource availability of each resource may change from planning period to planning period in order to better balance supply and demand in periods with irregular demand patterns. An activity may require a fractional number of time units of a resource.

Time

For some of the activities in each care process, there exist strict time limits that must be adhered to. For each care process, the order in which the activities are to be performed, is restricted. The goal of these restrictions is to ensure that a patient's activities are executed in the correct order. For example, before an MDT meeting takes place, all activities in a diagnosis phase must be finished on beforehand.

Decisions

The fundamental decision in the planning problem is to decide how many patients that are to be serviced from each queue in each time period of the planning period. These decisions in turn determine when and how much of each resource to allocate to each queue of patients.

4.3 The dynamic and uncertain context of the problem

The problem presented in this chapter is defined for a set of time periods that makes up one planning period. An important aspect of this thesis is to evaluate what happens to the system over multiple planning periods, that is its dynamics and the implications decisions at one point in time have for following planning periods. Assessing how the system behaves within one single planning period does not accentuate the system dynamics, nor yield results that produce managerial insights of significant value. Therefore, we describe the dynamics of the problem over multiple planning periods in the following. We also assess factors that contribute to uncertainty in the planning problem.

Each decision has an impact in the short run on the current planning period, but also on the following planning periods. It is often easy to detect the short term impact of a decision since we have a clear picture about the queuing situation on a day-to-day basis. But, revealing the impact of a decision in a longer run is more demanding. In fact, it might be impossible. From a decision is made to its effect becomes evident, the state of the system has most likely changed and queues may have grown.

To elaborate on the dynamic of the problem we refer to Figure 11. We assume that there are patients residing to the queues when the figure starts of a time P1. At time P1, the problem is assessed and a schedule for the first planning period is generated. Then, the schedule is implemented at the hospital from time P1 to P2. After the implementation period, one could of course just carry on and use the plans generated in the first planning period for time periods between P2 and P3. However, from P1 and P2, new information has been revealed e.g. the number of patients residing to each queue in the system may have decreased dramatically. Therefore, we suggest assessing the planning problem every time an implementation period ends and as a result update the schedule. By solving the problem again at P2, and not waiting until P3, planning is done more often and in closer proximity in time to the realization of the plans. The hope is that this enables a better control over uncertainties and balance between supply and demand.

In the implementation period uncertainties are accounted for. In the implementation period we evaluate how well the blueprint schedule for the planning period from the optimization problem, which is assessed in a deterministic setting, withstands uncertainties. We evaluate uncertainties associated with the number of patients referred to the system each time period, which treatment a patient demands after she has received her diagnosis - if she has cancer at all, how much time each patient uses at each activity and if a patient shows up to her activity or not. If these uncertainties were not present, we could solve the optimization model for a set of time periods, and be certain that this plan was adhered to. In a real life setting, plans are just plans, and not a recipe for how the future will turn out. It is these uncertainties that result in the need to plan more often.

If the planning period is longer than the implementation period, more stability is achieved by ensuring that the decisions made in the short-term are taken based on which events that might occur in the future. When the hospital has a longer outlook on the future, it enables them to more accurately foresee trends in the demand and react to these changes before it is too late. The hospital may observe where in the system the queues build up and act on these observations.

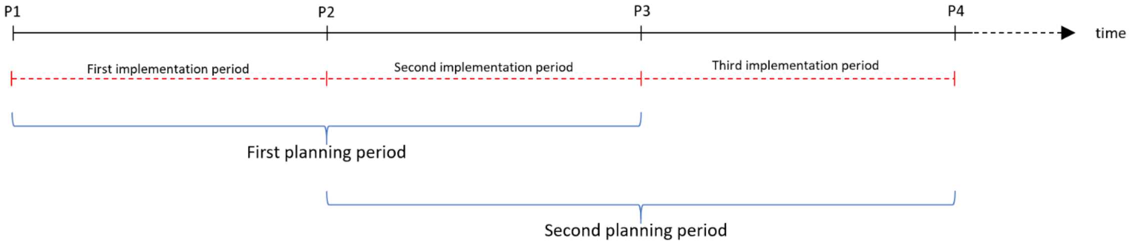

P1
P2
P3
P4
time
First implementation period
Second implementation period
Third implementation period
First planning period
Second planning period

Figure 11: The dynamics of the planning problem. P1, P2, P3 and P4 are points in time where planning is done. Here, the planning period is twice as a long as the implementation period. The planning problem presented in Section 4.2 is solved before a new implementation period starts.

5 Mathematical model

In this section, we present a mathematical model that seeks to represent the problem of scheduling patients in standardized care processes, as presented in Section 4.2. The mathematical model is a MIP model inspired by Hulshof et al. (2013) who model a tactical planning problem of resource allocation and patient admission planning. The mathematical model presented in this chapter is defined for one planning period only. However, it is of interest to assess the problem over multiple planning periods and its dynamics, as emphasized in Section 4.3. In order to evaluate the mathematical model over multiple planning periods, we introduce a rolling horizon approach. This approach is central to the solution of the model, found in Chapters 6 and 7.

In Section 5.1, the rolling horizon approach is described. Section 5.2 presents necessary notation in the mathematical model. Section 5.3 introduces the mathematical model. Lastly, in Section 5.4, modeling choices, assumptions and simplifications done when transcribing the physical problem to a mathematical optimization problem are presented.

5.1 Rolling horizon approach

We underscore that the mathematical model does not explicitly handle the uncertainties described in the problem description. The stochastic aspect of the problem is dealt with in the simulation model presented in Chapter 6. In the simulation model, uncertainty is replicated by drawing numbers from suitable probability distributions to imitate the stochastic, and more realistic behavior, of the problem. Since uncertainty is handled in the simulation model, stochasticity is separated from the mathematical model. Also, if stochasticity was to be included explicitly in the mathematical model, the degree of complexity would increase, both in terms of modeling, but also in terms of solution procedures. It is underlined that the simulation is not carried out by the hospital. We perform the simulation for evaluation purposes only and as a substitute for an actual implementation.

In a rolling horizon approach, the optimization model is solved iteratively. Each iteration consists of running the optimization model once and an implementation. When the optimization model is run, a schedule for all time periods in the current planning period is generated. We now have a plan ready for implementation. Instead of implementing the schedule for the entire planning period, only some of the time periods are implemented. The time periods of a planning period where the plan is implemented, are denoted as the implementation period. In the example in Figure 12, this corresponds to the first week of each iteration that is added to the implemented schedule. After the implementation, a new iteration starts. In the next iteration, the optimization model starts of where the last implementation left off. This ensures that the following iteration is coordinated with the previous iteration.

To clarify, we introduce an example. The example is illustrated in Figure 12. Here, the planning period is three weeks. First, the optimization model is run and a schedule for the entire first planning period, that is weeks 1, 2 and 3, is made. Here, the only time periods of the planning period that are implemented, in other words actually carried out by the hospital or simulated using a computer, are the time periods that make up week 1. Consequently, the implementation period in the example is one week. The plans from the first iteration for the time periods that are not implemented, those that constitute week 2 and 3, are kept as a reference for coming schedules and utilized as input in the optimization of the following planning periods. In the next iteration, the planning period is weeks 2, 3, and 4, and week 2 is added to the implemented schedule.

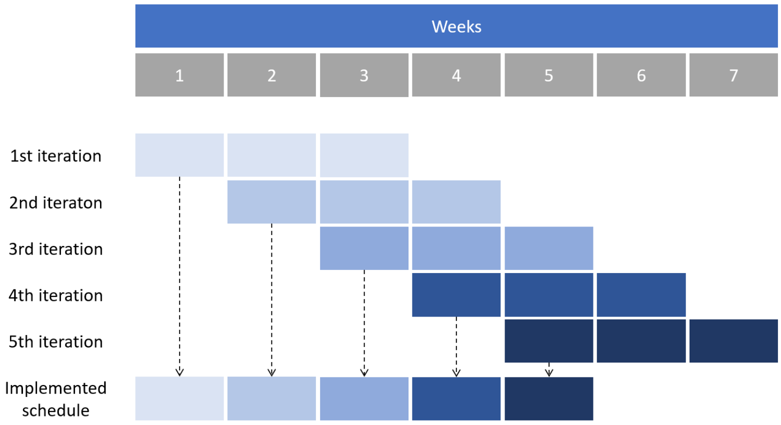


Figure 12: Rolling horizon pattern.

Our work differs notably from the work of (Hulshof et al., 2013) in several ways. Firstly, we are able to solve the optimization model for one planning period and, by utilizing the rolling horizon approach, assess how this solution influences the following planning periods. Secondly, to offer stability from planning period to planning period, we restrict the number of changes that can be made to a schedule compared to the schedule for a given queue and time period in the previous iteration. The model motivates a behavior where as much as possible of the schedule generated in one planning period is kept in the following planning periods. Thirdly, in our approach, constraints on time limits for certain queues are included to adhere to national guidelines on waiting times. Finally, we label each patient with an index m in order to explicitly track the total waiting time in its respective care process. We believe that by tracking the total waiting time for each patient instead of the individual waiting times at each queue, we are able to deliver swift health care services and prevent too long waiting times.

5.2 Notation

Sets are presented using an uppercase calligraphic font, indices are given as subscripts of lowercase letters, parameters are denoted using capital letters and variables are denoted using lowercase letters. Letters used for characterization of sets, parameters and variables are given as capital letters in superscript. For notation used in the mathematical model, see Tables 2, 3, 4 and 5 for indices, sets, parameters and variables, respectively.

Table 2: Indices used in the mathematical model.

Index	**Description**
i, j	Queues
t	Time periods
n	Time periods waited in a queue
m	Time periods waited in total
g	Patient care process
a	Activities
r	Resources

It is noted that n is incremented by one every time period a patient waits in a queue. The time period t a patient arrives in a queue, the counter n starts at zero. The index m keeps track of the sum of n for a patient, that is, it counts the total waiting time for each patient.

We let each queue be defined by an activity and care process, $j(g, a)$, meaning that the different queues a patient must wait in is defined by her care process g and its associated activities.

Table 3: Sets used in the mathematical model.

Set	Description	
$\mathcal{T}$	Set of time periods in the planning period	$t \in \mathcal{T}$
$\mathcal{T}^C$	Set of time periods in the current planning period that overlap with the time periods of the previous planning period	$t \in \mathcal{T}^C \subseteq \mathcal{T}$
$\mathcal{R}$	Set of resources	$r \in \mathcal{R}$
$\mathcal{J}$	Set of queues	$j \in \mathcal{J}$
$\mathcal{J}^T$	Set of the first queues in every treatment stage	$j \in \mathcal{J}^T \subseteq \mathcal{J}$
$\mathcal{J}^D$	Set of the last queues in every diagnosis stage	$j \in \mathcal{J}^D \subseteq \mathcal{J}$
$\mathcal{G}$	Set of patient care processes	$g \in \mathcal{G}$
$\mathcal{A}$	Set of activities	$a \in \mathcal{A}$
$\mathcal{E}^N$	Set of possible waiting periods in a queue	$n \in \mathcal{E}^N$
$\mathcal{E}^M$	Set of possible total waiting periods	$m \in \mathcal{E}^M$

Table 4: Parameters used in the mathematical model.

Parameter	Description
W_{jm}	Objective function weight of patients in queue j who have been waiting m time periods
D_{jt}	New demand in queue j in time period t
Q_{ij}	Possible transitions from queue i to j
L_{rt}	Resource capacity for resource r in time period t
H_{jr}	Resource demand for resource r for a patient in queue j
A_{jt}	The schedule for queue j in time period t from the previous planning period where $t \in \mathcal{T}^C$
M_{ij}	Minimum delay after being serviced from queue i before entering queue j
K_t^A	Maximum number of additional appointments to the schedule in time period t from the previous planning period
K_t^R	Maximum number of removals of appointments from the schedule in time period t from the previous planning period
F_j	Time limit for being serviced at queue j
E_{jnm}	The number of patients that have been waiting in n time periods in queue j and have waited in total m time periods in end of an implementation period
G_{jtm}	The number of patients that has spent m time periods in the system, but are not in a queue due to a delay, given by the parameter M_{ij}, and enter queue j in time period t in the next implementation period
P_{ij}	The share of patients that moves from the last queue $i \in \mathcal{J}^D$ in a diagnosis stage to the first queue $j \in \mathcal{J}^T$ in an associated treatment stage

Table 5: Variables used in the mathematical model.

Variable	Description
c_{jtnm}	The number of patients serviced from queue j who have been waiting n time periods in the queue, and m time periods in total by time period t
q_{jtnm}	The number of patients who have been waiting n time periods in queue j, and in total m time periods by time period t
b_{jt}	The number of patients serviced from queue j in time period t
u_{jt}^A	The number of additional appointments scheduled for an activity associated with queue j in time period t compared to the previous planning period
u_{jt}^R	The number of removed appointments for an activity associated with queue j in time period t compared to the previous planning period

5.3 Main model

$$\min z = \sum_{j \in \mathcal{J}} \sum_{t \in \mathcal{T}} \sum_{n \in \mathcal{E}^N} \sum_{m \in \mathcal{E}^M} W_{jm} q_{jtnm} \tag{1}$$

s.t.

$$q_{jt00} = D_{jt} + G_{jt0} + \sum_{i \in \mathcal{J} | t \geq M_{ij}} \sum_{n \in \mathcal{E}^N} Q_{ij} c_{i(t-M_{ij})n0} \qquad \forall j \in \mathcal{J}/\{\mathcal{J}^T\}, t \in \mathcal{T} \tag{2}$$

$$q_{jt0m} = G_{jtm} + \sum_{i \in \mathcal{J} | t \geq M_{ij}} \sum_{n \in \mathcal{E}^N} Q_{ij} c_{i(t-M_{ij})nm} \qquad \forall j \in \mathcal{J}/\{\mathcal{J}^T\}, t \in \mathcal{T}, m \in \mathcal{E}^M | m \geq 1 \tag{3}$$

$$q_{jt00} = D_{jt} + G_{jt0} + \sum_{i \in \mathcal{J}^D | t \geq M_{ij}} \sum_{n \in \mathcal{E}^N} P_{ij} c_{i(t-M_{ij})n0} \qquad \forall j \in \mathcal{J}^T, t \in \mathcal{T} \tag{4}$$

$$q_{jt0m} = G_{jtm} + \sum_{i \in \mathcal{J}^D | t \geq M_{ij}} \sum_{n \in \mathcal{E}^N} P_{ij} c_{i(t-M_{ij})nm} \qquad \forall j \in \mathcal{J}^T, t \in \mathcal{T}, m \in \mathcal{E}^M | m \geq 1 \tag{5}$$

$$q_{j0nm} = E_{jnm} \qquad \forall j \in \mathcal{J}, n \in \mathcal{E}^N | n \geq 1, m \in \mathcal{E}^M \tag{6}$$

$$q_{jtnm} = q_{j(t-1)(n-1)(m-1)} - c_{j(t-1)(n-1)(m-1)} \qquad \forall j \in \mathcal{J}, t \in \mathcal{T} | t \geq 1, n \in \mathcal{E}^N | n \geq 1, m \in \mathcal{E}^M | m \geq 1 \tag{7}$$

$$c_{jtnm} \leq q_{jtnm} \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, n \in \mathcal{E}^N, m \in \mathcal{E}^M \tag{8}$$

$$b_{jt} = \sum_{n \in \mathcal{E}^N} \sum_{m \in \mathcal{E}^M} c_{jtnm} \qquad \forall j \in \mathcal{J}, t \in \mathcal{T} \tag{9}$$

$$\sum_{j \in \mathcal{J}} H_{jr} b_{jt} \leq L_{rt} \qquad \forall r \in \mathcal{R}, t \in \mathcal{T} \tag{10}$$

$$u^A_{jt} - u^R_{jt} = b_{jt} - A_{jt} \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}^C \tag{11}$$

$$\sum_{j \in \mathcal{J}} u^A_{jt} \leq K^A_t \qquad \forall t \in \mathcal{T} \tag{12}$$

$$\sum_{j \in \mathcal{J}} u^R_{jt} \leq K^R_t \qquad \forall t \in \mathcal{T} \tag{13}$$

$$q_{jtnm} = 0 \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, n \in \mathcal{E}^N, m \in \mathcal{E}^M | m > F_j \tag{14}$$

$$b_{jt} \in \mathbb{Z}_{\geq 0} \qquad \forall j \in \mathcal{J}, t \in \mathcal{T} \tag{15}$$

$$c_{jtnm} \geq 0 \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, n \in \mathcal{E}^N, m \in \mathcal{E}^M, m \geq n \tag{16}$$

$$q_{jtnm} \geq 0 \qquad \forall j \in \mathcal{J}, t \in \mathcal{T}, n \in \mathcal{E}^N, m \in \mathcal{E}^M, m \geq n \tag{17}$$

$$u^A_{jt} \geq 0 \qquad \forall j \in \mathcal{J}, t \in \mathcal{T} \tag{18}$$

$$u^R_{jt} \geq 0 \qquad \forall j \in \mathcal{J}, t \in \mathcal{T} \tag{19}$$

Objective function

The objective function minimizes a weighted sum of patients waiting in a queue for an activity in their care process over the planning period. The hospital is free to assign values to the objective function weights based on their own preferences. An alternative technique for assigning values to the weights is elaborated on in Appendix B.

Constraints

Constraints (2) to (5) secure that patients are assigned the value 0 for the index n for the set of variables q_{jtnm} when they arrive at a queue. Constraints (2) define the initial queuing situation for patients that have arrived to the system during the most recent time period, that is $m = 0$. Constraints (3) serve the same purpose as the preceding set of constraints, but handle patients that have been waited in the system for more than one time period, that is for patients with $m > 0$. Constraints (4) and (5) ensure the right share of patients to be transferred from the diagnosis stage to the treatment stage. Here, each possible transition from a diagnosis stage to a treatment stage is assigned a certain probability of happening. In a deterministic setting, this is equivalent to the transition rate from a specific diagnosis stage to a specific treatment stage. In addition, constraints (4) handle external patients that are transferred directly into the treatment stage, and patients that not have been accumulated any waiting time throughout the system.

Constraints (6) initialize the queues in the first time period in the current planning period by importing the ending queue situation from the previous iteration of the model. Constraints (7) updates the queues from one time period to the next by removing patients that are serviced and incrementing indices for patients that are still in the queue.

Constraints (8) limit the number of patients in a queue that possibly can be treated to the number of patients that are in that queue in the relevant time period. They also ensure that not more than the number of planned patients are treated in the given time period. Constraints (9) aggregate the variables c_{jtnm} into the set of variables b_{jt}. These new variables denote the total number of patients serviced from queue j in time period t. Constraints (10) hinder overexploitation of resources, by stating that patients can not consume more resources than the amount of resources that is available in the relevant time period.

Constraints (11) link the current schedule to the schedule generated for the previous planning period and counts the deviations between the two schedules for all time periods where there is overlap (Figure 12). Constraints (12) and (13) represent the upper bounds of changes, additions and removals, respectively, that can be made to overlapping time periods between the current and former planning period.

The constraints (14) ensure that no patients are waiting longer than the time limit for reaching the queue j. Note that not all queues have an associated time limit, only those in the set F_j. In the context of CCPs, these time limits are only present in the diagnosis stage. This set of constraints also ensures that there is an upper bound of time periods a patient can stay in the system in total.

Constraints (15) to (19) are restrictions on the variables in the problem. Constraints (15) are integer requirements on the set of variables b_{jt}. These constraints guarantee that the total number of patients serviced from every queue in every time period, constraints (9), is an integer number. Constraints (16) and (17) ensure that variables of type c_{jtnm} and q_{jtnm} are only defined for queues with valid indices, that is as long as the total waiting time m is larger than or equal to the waiting time n in a single queue. Constraints (18) and (19) are non-negativity constraints for variables u^A_{jt} and u^B_{jt}.

5.4 Modeling choices and assumptions

Modeling choices

A key aspect of the mathematical model are the queues. Each patient follows a care process g that is made up of a diagnosis stage and a treatment stage. In each stage, there are a set of activities that must be undertaken. Each patient must queue for each activity and there is one queue for each care process for each activity.

To illustrate how the patients advance in a queue, we would like to introduce an example which is visualized in Figure 13. In this example, we have limited the maximum time periods a patient can wait in a queue to three time periods. For the case of simplicity, we only assess one queue j and assume all patients to have spent the same total number of time periods m in the system.

The prioritization of patients are taken care of in the process of assigning values to the objective function weights.

In the example, the beginning situation in queue j at time $t = 1$ is that one new patient enters the queue, two patients have waited one time period in this queue already, one patient has waited two periods and two patients have waited three time periods. When running the model, it is decided that two patients from this queue are serviced ($c_{j13m} = 2$). In this case, the model chooses to service the two patients that have waited the longest. These two patients are then sent two the next queue in their care process. It is noted that the model is not strictly obliged to choose the patients that have waited the longest in the queue. In time period $t = 2$, the patients that was not serviced in time period $t = 1$ are assigned new positions in the queue, with n incremented by one. Three new patients arrive in the queue this time period. Three patients are serviced, of whom two have waited two time periods and one have waited three time periods. In time period $t = 3$, the patients that were not serviced in time period $t = 2$ are assigned new positions in the queue with n incremented by one. One new patient arrives in the queue in this time period. The example ends here, but the process will in practice continue for all t in $\mathcal{T}$.

It is noted that although all the variables in this example hold integer values, this is not a condition that must be held and the variables q_{jtnm} and c_{jtnm} may take fractional values in the optimization model.

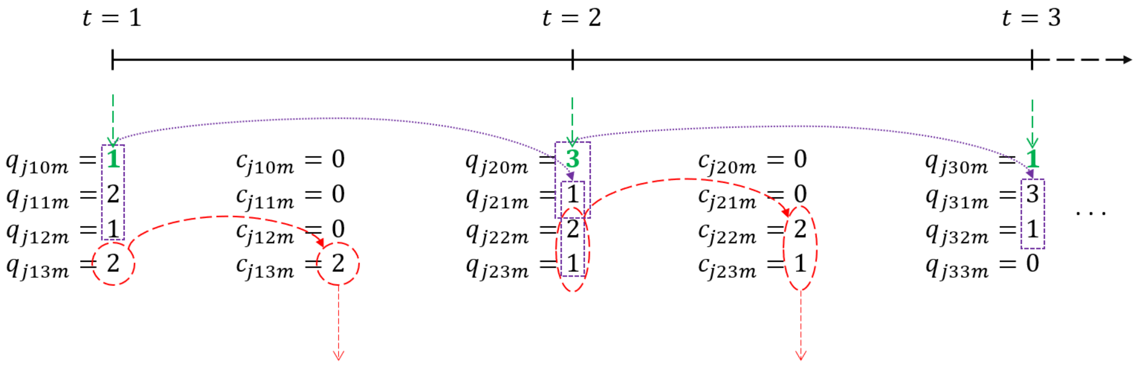


Figure 13: Queue dynamics. In order to simplify the figure, we depict one queue j and assume that all patients have the same m.

Assumptions in the mathematical model

Some assumptions and simplifications must be made in the transition from the actual problem to a mathematical description of it. The intention of the assumptions is to delineate the problem, but at the same time preserve the core aspects of it. Underlying assumptions in our modeling approach are the following:

Table 6: Assumptions in the mathematical model.

	Description
Assumption 1	In the mathematical model, patient arrivals, resource requirements, resource capacities and the minimum delay between queues are considered to be deterministic and known.
Assumption 2	If illness is not detected after the diagnosis stage, the patient is discharged.
Assumption 3	A patient can not have stayed in a queue longer than she has stayed in the system in total, implying that $n \leq m$.
Assumption 4	The parameters E_{jnm}, G_{jtm} and A_{jt} are input data to the mathematical model from the previous planning period.
Assumption 5	No patients are serviced during weekends.
Assumption 6	Resource capacity L_{rt} for resource r in time period t is not transferable to other time periods. Unused capacity is considered lost.
Assumption 7	Patients in the population get ill independently of each other.

6 Simulation model

The problem of scheduling patients in a standardized care process is complex and dynamic, as elaborated on in Section 4.3. Also, in a hospital setting, human factors contribute to uncertainties. The optimization model formulated in Chapter 5 does not explicitly handle the uncertainties introduced in the problem description. The optimization model presented there is solved using exclusively deterministic input values.

However, the stochasticity should preferably be taken into account. We do not introduce stochasticity in the solution of the optimization model, but rather test how the solution from the optimization model performs in a stochastic setting using simulation. It is important to emphasize that the proposed simulation model is not something the hospital is going to implement or put to use, it is rather a tool designed with respect to this thesis for evaluating the schedules generated by the optimization model. Furthermore, it is noted that this chapter only provides a description of the simulation model itself. The interaction between the optimization model and the simulation model in a scheduling framework is presented in Chapter 7.

In this chapter, the simulation model is presented. First, in Section 6.1, we provide a description of the system components of the system we simulate. Next, in Section 6.2, the configuration of the simulation model is provided. Here, simulation model assumptions, a fixed-increment simulation model and a discrete event simulation model are presented.

6.1 System components

To present the components of the system, we use the framework for describing a simulation model proposed by Hillier and Lieberman (2015). The system that is to be simulated is a hospital department (in our case: a department of gynecological cancer) that schedules patients undergoing a standardized care process.

The state of the system

The state of the system is defined by information available on the patients that are queuing for an activity and the patients that are being serviced at any time. The model provides information on how many patients that are queuing in the different queues, how long each patient has waited in the queue it belongs to at the moment and how long each patient has stayed in the system in total. The possible states of the system are built up by all combinations of patients queuing and being serviced that adhere to the restrictions of the problem, presented in Chapter 4.

Possible events

The simulation distinguishes between two events: patients may either arrive at a queue or be serviced. When an event occurs, the simulation model reacts and the state of the system is updated correspondingly.

Simulation clock and method for randomly generating events

To perform this simulation, we need two simulation clocks. The first clock is incremented once every time period, for example every day. The simulation spans multiple time periods, and the simulation must update when we travel from one time period to the next. Consequently, this method for updating the simulation clock is a fixed-time increment (Hillier and Lieberman, 2015). In this thesis, the set of time periods this clock simulates is denoted the simulation period. The simulation period is comprised by a set of implementation periods, as introduced in Chapter 5.

The second clock is incremented every time an event occurs in a time period, that be a patient arriving at a queue or a patient being serviced. This type of advancing the simulation clock is by Hillier and Lieberman (2015) denoted a next-event increment. The next event that occurs is determined based on which event the simulation clock triggers first and is based on a discrete event simulation. The probability distributions utilized to generate the random events in this thesis are revealed in the computational study in Section 8.1.

A formula for identifying state transitions

When an event occurs the simulation clock and the system is updated in response to it. Patients are assigned a position in a queue when they arrive at a queue. When a patient has been serviced, she is sent to the next queue, possibly with a delay. This delay is added for medical reasons, for example if a specific treatment forces the patient to wait a certain number of time periods before she proceeds in her care process. A more extensive explanation of the queue dynamics in the simulation is provided in Sections 6.2.2 and 6.2.3.

6.2 Configuring the simulation model

In Section 6.2.1, we state assumptions necessary to perform the simulation. These assumptions are closely related to the uncertainties presented in the problem description in Chapter 4. The simulation model consists of two simulations that serve two different purposes. First, we have the fixed-increment simulation model, which manages the invocation of new time periods, explained in Section 6.2.2. Then, the discrete event simulation, described in Section 6.2.3, is run. This simulation manages the flow of patients through the different queues within a time period. The two models are tightly connected and work together in a loop during the whole simulation period.

6.2.1 Assumptions in the simulation model

The simulation, in contrast to the optimization model, handles uncertainty explicitly. It is assumed that new patients are referred to the system every time period in a random fashion in accordance with some fixed probability distribution. The new patients enter the system in the beginning of every time period, and from this point in time and onward they may be scheduled to undergo activities. Patients may be referred to the hospital in each time period the hospital admits patients. It is assumed that the population we draw patients from is infinitely large, i.e., the rate of referrals is unaffected by how many patients that already have been drawn.

For each patient arrival process, there exists a probability for no-show. A no-show occurs when a patient does not show up for an appointment. The capacity that was reserved for this appointment can not be used by other patients and is considered lost. The patient that did not show up is not eligible for service the current day. Which position the patient is given after a no-show, is dependent on the scheduling policy. A patient not showing up, is moved to the last spot of the queue associated with the activity she did not show up to on time. Different scheduling policies are discussed and evaluated in Section 9.3.2.

The amount of time, and consequently amount of resources, a patient requires at an activity when being serviced, varies. In the simulation, we assume that each activity has a service time of random length in accord with some fixed probability distribution.

The simulation takes into account uncertainty linked to which treatment path a patient is assigned to. When a patient is finished with her diagnosis stage, it is decided whether or not she has to undergo treatment, and in the case of treatment, which treatment she will receive. Which, if any, treatment a patient receives after diagnosis, is in the simulation decided upon based on a probability distribution. This probability distribution is based on empirical data obtained from the relevant hospital department.

We also assume that patients may only leave the system at two points. First, if the diagnosis stage reveals that a patient has no cancer illness, there is no need for the patient to receive any treatment. As a consequence, patients may leave the system after the diagnosis stage. Secondly, patients may leave the system after they have received their treatment. The problem is only defined for the part of the care process up until the treatment. Therefore, follow-up and evaluation is kept out of the simulation.

6.2.2 Fixed-increment simulation model

The optimization model is run prior to the simulation. How the optimization model interacts with the simulation is described in detail in Chapter 7. The optimization model provides an upper bound, in the form of a scheduled capacity, for how many patients that can be treated from a single queue in each time period during the simulation period. After the optimization model has generated the schedules, the fixed-time increment model, which is visualized in Figure 14, begins its simulation by invoking a new time period.

The start of a new time period entails several routines. First, the time period index t is incremented by one. Then, the new arriving patients to the system are generated. This is done by drawing a number of patients from a probability distribution for each of the care processes. This implies that patients arrive in batches to the different care processes at the beginning of each time period. Thereafter, all the queues are updated with arrivals of patients that have undergone activities that have lasted over multiple time periods, patients that for other reasons have been delayed (e.g. medically justified delays) or patients that have had a no-show. Furthermore, all the patients residing in a queue get their waiting time incremented.

When the mentioned updates are done, the fixed- time increment-model hands over the system to the discrete event model in order to manage the flow of patients through the different queues in the current time period. When the discrete event simulation is done, it passes the system state back to the fixed-increment simulation model. Based on the number of defined simulation periods, the simulation either terminates or invokes another time period. If the simulation period is over at this point in time, the simulation stops. Otherwise, a new time period is invoked and the previous steps are repeated in a loop until the simulation period eventually reaches its end. This loop is elaborated on in Chapter 7.

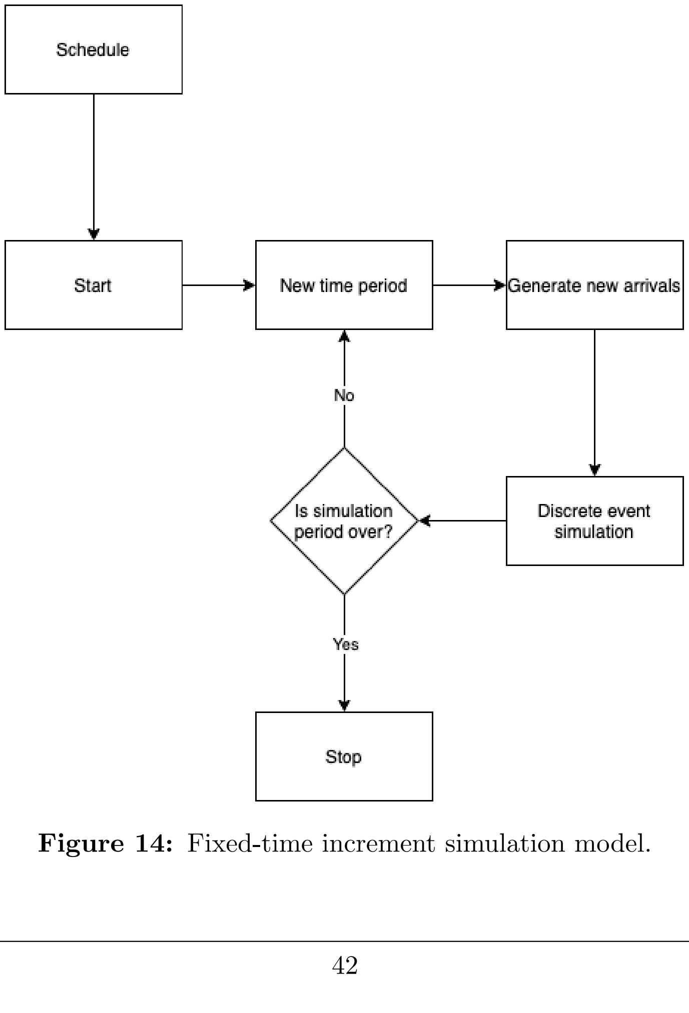


Figure 14: Fixed-time increment simulation model.

6.2.3 Discrete event simulation model

The discrete event simulation model is illustrated in Figure 15. This model functions as an inner module of the fixed-time increment model, managing the flow and routing of patients through their care processes during a single time period. The simulation inputs the system state by the start of the time period, and outputs the system state at the end of the same time period. Within a time period the model reacts to events that occur. The next event in the system within a time period is decided on by identifying the service event or arrival that lies closest in time.

If the next event is a service event, the queue containing the patient labeled with the next service is chosen. The model chooses the patient that is to be serviced from the queue, based on a scheduling policy. When a patient is chosen to be serviced from a queue, there exists a probability of no-show. If this actualizes, the patient may not be serviced from the queue the current time period and has to be rescheduled to the next time period before she can be selected for service again. On the other hand, if the patient shows up and leaves the queue for service, the patient's label of number of days in the queue, n, is set to 0 and a service time is drawn from a probability distribution. If the patient is not to be treated after the diagnosis stage or has reached the end of the care process, she is discharged at this point and leaves the system. If the patient is not discharged, the next queue is updated with her time of arrival. This time of arrival is dependent upon the generated service time and the expected delay associated with the undergone activity. It is important to notice that the patient is not in the next queue before this time has passed. Thus, when the service time plus delay has passed and the simulation clock is incremented to the corresponding time, the arrival into the next queue takes place. After the service of a patient, the queue checks if there are enough resources in order to conduct more services. If there are not enough resources, the queue is not eligible for any more service events in the current time period.

If the next event is an arrival, the patient has been serviced at another queue earlier in the same time period. If a patient arrives at an empty queue, and there is idle capacity allocated to serve this queue, the service of this patient in this queue is initialized immediately. If that is not the case, the patient must queue for the activity until she holds a spot in the queue that is serviced. For every time period the patient is enqueued, the values for n and m are incremented.

The discrete event simulation model repeats itself until there are no more events that may happen during the current time period. The model knows that there are no more events to happen when it can not find new arrival or departure events. This can be caused by no more available capacity in the queues, or the resources have reached their capacity, thus meaning no service events may take place. No service events further entails no possibilities of arrival events. Furthermore, a situation where there are no patients in any of the queues also leads to no events happening. If no new events are found, the model stops and sends the system state back to the fixed-time increment model.

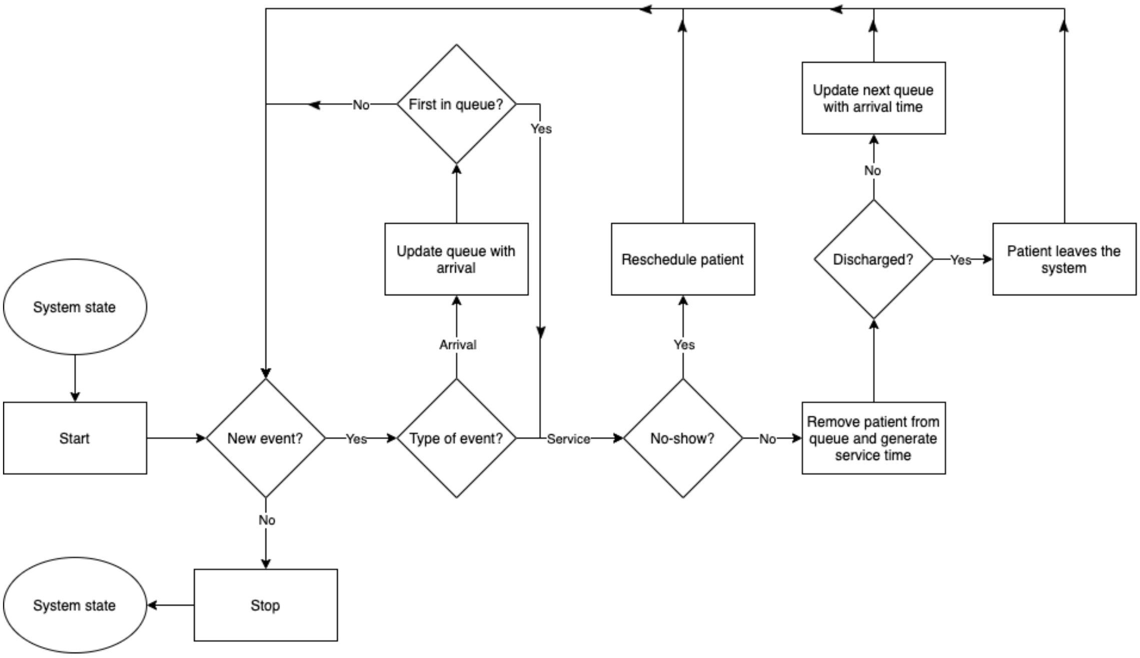


Figure 15: Discrete event simulation model.

7 Scheduling framework

In this chapter, a combined optimization and simulation framework is proposed. The framework is inspired by the approach proposed by Addis et al. (2015). The motivation behind the choice of framework is discussed in Section 7.1. Thereafter, the framework and its components are explained in Section 7.2. Results from running the scheduling framework make up the basis for the computational study provided in Chapter 9.

7.1 Motivation for design of framework

As introduced in Chapter 4, the goal of this thesis is to generate schedules where the number of patients that are to be serviced from each queue in each time period in the planning period is decided on. The optimization model presented in Chapter 5 is deterministic, even though the hospital environment is uncertain.

An important reason for designing the framework presented in this chapter is that there by this point is no opportunity to test the schedules in a real-life setting. Therefore, it is necessary to create an environment where the system can be evaluated and surveyed in a realistic setting over multiple planning periods. One purpose of the suggested framework is hence to mimic how the generated schedules perform when they are implemented in a real environment. However, in this thesis the implementation in a real environment is substituted with a simulation. Simulations pose a quick and relatively cheap way of testing and assessing schedules before they eventually are implemented in real-life. This means that it is not intended for the hospital to implement the scheduling framework. This is because the framework is only used to evaluate the decision tool we propose for the hospital, namely the optimization model.

7.2 Scheduling framework

One loop in the framework, visualized in Figure 16, corresponds to one iteration, as defined in Figure 12. Each iteration consists of three steps invoked in a consecutive order: optimization (step 1), simulation (step 2) and updating the optimization model (step 3). The output of each step feeds the next step in the current iteration, indicated by the arrows. Square boxes generate results, the ellipses illustrate results and diamonds delineate aggregation of inputs that are to be passed into the optimization model. The framework is iteratively executed a number of times specified by the system user.

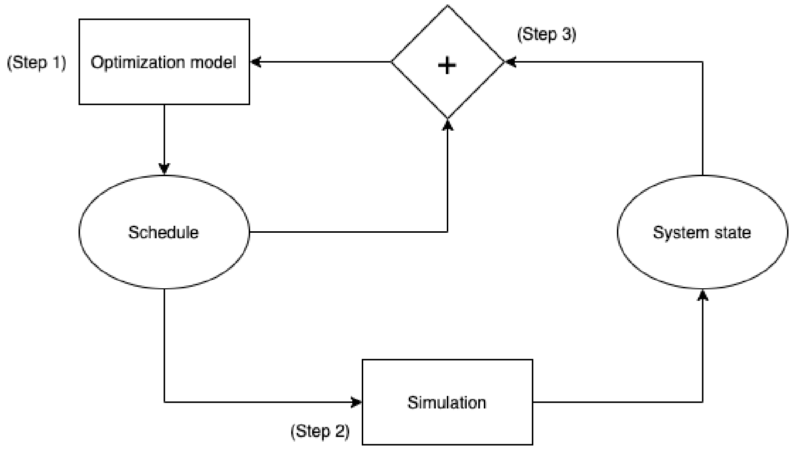


Figure 16: Structure of the scheduling framework

Each iteration of the loop begins with running the optimization model. This step returns a schedule for the planning period, but only the schedule for the implementation period is forwarded to the next step. The schedules for time periods in the planning period that are not part of the implementation period are passed on to the next iteration. These schedules serve as input to a look-ahead model that will help generating more accurate schedules. Then, the simulation step simulates how well the plan withstand uncertainty by imitating a real-world implementation of the schedule for the implementation period. The simulation step returns the state of the system. Finally, the system state and the remainder of the schedule from step 1 is fed into the optimization model, ending one iteration of the framework.

The framework starts out as an empty system. This means that no patients are in the system, and no resources are occupied initially. This causes the initial observations of waiting times to tend to be smaller than if the system was in a steady-state. Therefore, an initial warm-up period is required. A steady-state is obtained when the balance between inflow and outflow of patients are stable. After the warm-up period is complete, one can start recording information from the system. Since the simulation model takes uncertainty into account, results vary slightly if the framework is executed more than once. Therefore, the framework should be run several times and reporting should preferably be done on the background of averages from the runs.

Simulation framework classification

Figueira and Almada-Lobo (2014) introduce the classification of approaches to solution generation and solution evaluation for simulation-optimization problems. In a solution generation approach, the main objective is to contribute to the generation of a solution by using simulation, whereas a solution evaluation uses simulation to evaluate the performance of various solutions (Figueira and Almada-Lobo, 2014).

Classifying our scheduling framework is not straightforward. On one side, we may look at the framework as a solution generation procedure, i.e. schedules are being produced by the optimization model and forwarded to the simulation model in order to create a solution. On the other hand, our scheduling framework may also be interpreted to take a simulation evaluation approach. This is because the framework also uses the simulation to evaluate the performance of the different solutions generated by the optimization model. Therefore, the scheduling framework proposed in this thesis may be classified as a mix of the two, according to the taxonomy of Figueira and Almada-Lobo (2014).

7.2.1 The optimization model (Step 1)

The optimization model generates schedules according to an estimated demand. This demand is based on two factors: the expected number of arrivals of new patients and the number patients already in a care process by the end of the previous iteration. Given this, the optimization model generates a plan for the next $\mathcal{T}$ time periods. The plan encompasses how many patients that are to be treated in the different time periods in the planning period, as well as which corresponding resources that needs to be available in which of the time periods.

Although a plan for the entire planning period $\mathcal{T}$ is generated, it is emphasized that only the time periods making up the implementation period is simulated in step 2. The remainder of the plan is passed on to step 3, where it is retained as a reference for the next iteration. This forward feeding of information onto the next iteration is central to the rolling horizon approach presented in Chapter 5. In addition to reducing the patient waiting times through the system, the optimization model also restricts the number of deviations in a schedule in one iteration compared to the schedule generated in the previous iteration.

7.2.2 Simulation (Step 2)

The purpose of step 2 is to use the generated schedule from the previous step to simulate how the solution functions in a simulated environment aiming at mirroring the reality. This step behaves according to the description provided in Chapter 6.

In each loop of the scheduling framework, the simulation starts out from where it was at the end of the previous loop. The only addition is the new schedule from the previous step in the current scheduling framework loop. In each time period of the implementation period, the number of new arriving patients to each care process are drawn from a probability distribution. The simulation model also takes into account uncertainties in service times, which treatment a patient shall receive and no-shows. The simulation is conducted for as many time periods as the implementation period's indicates. At the end of the simulation step, the system state is recorded and passed on to the next step.

7.2.3 Updating the optimization model (Step 3)

In the last step of the scheduling framework, the portion of the planning period generated in the current loop that was not implemented and the current system state are collected. The remaining schedule is used in the following loop of the scheduling framework in order to provide predictability in the planning process for the hospital, by reducing the number of allowed alterations of planned appointments. Step 3 marks the end of one iteration of the scheduling framework.

8 Input data

The input data required to perform the computational study in Chapter 9 is presented in this chapter. In Section 8.1, we present input data for the optimization model and in Section 8.2, we present data specifically for the simulation model. The two models demand different input since their tasks differ. The optimization model seeks to generate schedules for patient admission and resources during the planning period, whereas the simulation model aims at replicating a realization of these schedules in order to evaluate their performance. The data is inspired by Oslo University Hospital. The collection of the data was carried out in collaboration with the staff at Oslo University Hospital. From this collaboration, there are two types of data sources. The first data source is spreadsheets crafted for this thesis. The other source of data is information obtained through conversations. The data obtained from these sources is information about patient demand, resources, the order of activities in the different cancer care pathways, how much resource each activity demands and resource capacities. Table 9 lists some of the parameters and set sizes that have been collected, and further elaborations on the data are performed in the following.

8.1 Input data for the optimization model

This section presents the input data that is applied when solving the optimization model. The data that is used is inspired by information gathered from the Department of Gynecological Cancer at Oslo University Hospital and serves as a foundation for the computational study. The data aims at mirroring the situation at Oslo University Hospital, but in some cases adjustments have been made to make the environment fit into the presented model. These adjustments have been made because the hospital environment is complex and it is difficult to model every nuance of the problem and account for all the details. The adjustments and assumptions for each parameter are discussed in the following sections. In Section 8.1.1, we elaborate on the different types of data collected and look into how the obtained data is translated into the parameters we have defined in Chapter 5.

8.1.1 The different types of data collected and generated

Patient demand

The demand for each cancer pathway is based on a yearly arrival rate (Figure 17). In order to extract the daily demand on the opening days of the hospital, the yearly amount of arrivals have been averaged over the days in a year, excluding Saturdays and Sundays. Thus, the amount of expected arrivals on each opening day is assumed to be the same. Additionally, since the number of daily arrivals is based on an average value, it may take fractional values as we see in Table 9. This means that the problem does not consider seasonality. Furthermore, it is assumed that new patients that arrive to the system only may enqueue in the first queue in a care process. This implicates that no patients may arrive in later queues or stages of a care process, without going through the previous activities leading to this specific queue. In other words, the aforementioned external patients are neglected.

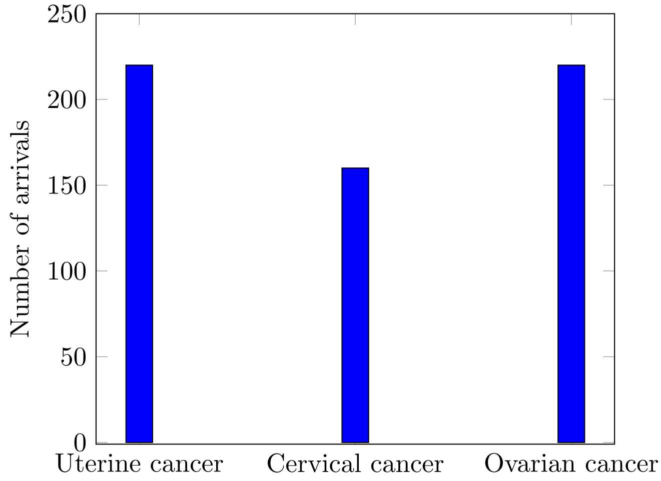


Figure 17: Number of arrivals in the different cancer groups per year.

Furthermore, based on the number of patients that are enrolled in the different treatment stages and the total number of arriving patients in the different care processes, an estimate of the parameter P_{ij} is made. This set of parameters denotes the share of patients that is routed between a specific diagnosis stage and the different treatment stages.

The care processes, activities and queues

As described in Section 2.2, there are 28 cancer care pathways that are standardized in Norway. The gynecological cancers mainly treated at the Department of Gynecological Cancer at Oslo University Hospital make up three of these. The hospital differentiates between the diagnosis and treatment stage in a cancer care pathway. The diagnosis stage usually consists of 4 or 5 activities, whereas the treatment stage may be somewhat shorter or longer, ranging between 1 and 8 activities.

Table 7 and 8 show the sequence of activities in the different diagnosis and treatment stages associated with the three gynecological care pathways. The abbreviations OC and GE mean outpatient clinic and gynecological examination, respectively. Since a care pathway may have several different treatment paths associated with it, each sequence of activities is labeled with the cancer type and a number, identifying the associated care pathway and the treatment path. This means that a patient following the diagnosis activities of a uterine care pathway, either enrolls in the Uterine 1 or Uterine 2 treatment path, or is discharged after the diagnosis.

We see that the diagnosis activities for the different care pathways are much the same and of the same length, whereas the treatment activities are of various lengths and activities.

Table 7: Diagnosis activities.

Cancer	**Activities**
Uterine	Referral → GE → Biopsy → CT
Cervical	Referral → OC → MRI → CT → GE
Ovarian	Referral → Biopsy → OC → MDT

Table 8: Treatment activities.

Cancer	Activities
Uterine 1	OC → GE → Surgery
Uterine 2	OC → GE → Surgery → Chemoterapy
Cervical 1	Surgery
Cervical 2	Surgery → MDT → Radiotherapy
Cervical 3	CT → Chemoterapy → Radiotherapy → Blood sample → Radiotherapy → Brachyterapy → Chemoterapy → MRI
Ovarian 1	OC → Chemoterapy → Surgery → OC → Chemoterapy
Ovarian 2	Surgery → OC → Chemoterapy

Resources

Each activity in the cancer care pathway is associated with at set of resources. The resources are dedicated to the associated activity during the service time of the patient. The number of resources per activity range from 2 to 7. In the case of a gynecological examination, three resources are needed, i.e. a gynecologist, a nurse and a room at the outpatient clinic. Whereas during surgery, several physicians, tools and nurses are required. Each resource has a daily capacity, and in this thesis they are assumed to be the same for every day in the planning period, except during weekends where all resource capacities are assumed to be zero. This way, no patients are serviced during this time. Furthermore, it is assumed that the resources cannot exceed their capacity, and overtime is therefore not possible. Resource capacities are given as a number of time units available each day, and an activity demands a given amount of time units, which may be fractional, of one or more resources in order to be performed. An overview of each of the activities and the associated resources and the capacities for the resources, is given in Appendix C.

Deciding the size of the possible number of waiting periods

The sets $\mathcal{E}^N$ and $\mathcal{E}^M$ determine the number of possible waiting periods in respectively a queue and the whole system. One wish to set these sets as low as possible in order to reduce the problem size, but still ensure that they are large enough to capture all possible queue positions that may develop when running the model.

The objective function weights

As stated in Chapter 5, the objective function minimizes a weighted sum of patients waiting in a queue for an activity in their care process over a set of time periods. The objective weights are functions of the j and m indices for the associated q variables. The objective function is defined in this manner to enable the hospital to prioritize different care pathways. The objective function weights that are used in the optimization model were not set in collaboration with Oslo University Hospital and are decided to be independent of the j index. This is because all the pathways and the associated queues are controlled by the same time limits, making the situation equal for each of the care pathways. Therefore, it was not relevant to prioritize between the different care pathways, and hence the objective weights are not dependent on the queues. Furthermore, we assigned the weights such that patients who have been waiting more time periods in the system are prioritized over those who have waited less. The objective weights are dependent on the m index of the associated q variable in the objective function, such that $W_{jm} = m$, which results in the objective function seen in Equation (20). In this manner, the more time periods a patient waits in the system, the higher is her priority.

$$\min z = \sum_{j \in \mathcal{J}} \sum_{t \in \mathcal{T}} \sum_{n \in \mathcal{E}^N} \sum_{m \in \mathcal{E}^M} m q_{jtnm} \tag{20}$$

Time periods and planning period

One time period t is assumed to equal one day. The length of the planning period is not determined by Oslo University Hospital, but is assumed to be 2 weeks. The longer the planning period, the bigger and more complex the problem gets. A long planning period may yield more predictability for both the hospital and the patients, since the schedules are known a longer time in advance. This predictability comes with a cost of flexibility, since the model tries to prevent changes in the schedule. On the other hand, a short planning period serves as a contrast to the above described. It results in a less complex problem, less predictability, but more flexibility. The effects of changing the length of the planning period are evaluated in Chapter 9.

Other parameters

The parameters defining the number of days after an activity is undergone until the patient enters the next queue, M_{ij}, were defined in collaboration with Oslo University Hospital. The values in the diagnosis phase are usually below 1 day, while the treatment stage may have activities with values spanning several days. For example, after a patient is done with surgery, she needs 4 days of recovery until she may be eligible to be serviced at the next queue.

The parameter F_j denotes the time limits for reaching a certain queue. These parameters are settled by the Norwegian Directorate of Health, as discussed in Section 2.2. The time limits for the different phases of the diagnosis are presented in Table 9. The values define time from referral to first attendance at the hospital, first attendance until diagnosis phase is completed and from the completed diagnosis phase until start of treatment, respectively. In this thesis, we assume that these values are the same for all patients, independent of which cancer diagnosis they may have. Each time limit is connected to the different queues that represent the events described above. The remaining queues are also defined with a time limit, but these limits are set sufficiently high in order to avoid the possibility of patients residing in a queue in an infinite amount of time periods.

K_t^A and K_t^R, which control the number of allowed deviations from previously planned schedules, are not set in collaboration with Oslo University Hospital. A high value leads to strict schedules, while low values leads to the opposite. The daily number of disruptions that is allowed, is dependent on the total weekly demand. Thus, more patient arrivals lead to a greater acceptance for changing the schedules, and opposite, a decrease in patient arrivals leads to less opportunities to change the schedules.

Parameters not discussed

The remaining parameters not discussed in this section are left out because they are derived from other sources than Oslo University Hospital or discussed earlier in this thesis. That is parameters such as A_{jt}, which denote the previously generated schedule. Additionally, the E_{jnm} and G_{jtm} parameters are not discussed. The E_{jnm} parameters define the number of patients in a queue by the end of the previous implementation period, while the G_{jtm} parameters denote the number of patients that are not enqueued, but are recovering from a previously undergone activity and is scheduled to enqueue in the next implementation period. These parameters are not defined because they are automatically generated when the scheduling framework is run.

Table 9: Overview of input data in the optimization model.

Data	Description	Used values for testing
$\mathcal{G}$	Number of patient processes	3
$\mathcal{T}$	Number of time periods in the planning period	14
$\mathcal{T}^C$	Number of time periods in the current planning period that overlap with the time periods of the previous planning period	7
$\mathcal{R}$	Number of resources	14
$\mathcal{J}$	Number of queues	52
$\mathcal{G}$	Number of patient care processes	3
$\mathcal{A}$	Number of activities	12
D_{jt}	Demand per time period for the three care processes	0.8, 0.6, 0.8
F_j	Time limits for being serviced at queue j	6, 22, 36

8.2 Input data for the simulation model

In this section the input data for the simulation model is discussed. First, in Section 8.2.1 the different probability distributions in the simulation model are elaborated on. Next, in Section 8.2.2, the scheduling policy is described. Finally, in Section 8.2.3, the warm-up period of the simulation is described.

8.2.1 The probability distributions

As covered in Chapter 6, there are four main sources of uncertainties involved in the simulation model. These sources are patient arrivals, service time, no-shows and the transition from the diagnosis stage to the treatment stage or a discharge. Associated with each of these four sources of uncertainty there is a an underlying probability distribution. The assumed distributions are summarized in Table 10.

Table 10: Probability distributions used in the simulation.

Event	Distribution
Patient arrivals	Poisson distribution
Service time	Empirical discrete distribution
Transition from diagnosis to treatment	Empirical discrete distribution
No-show	Bernoulli distribution

Patient arrivals

The number of patients that arrive to the system in the different care processes are drawn from a Poisson distribution. This may be done based on the assumption that patients get ill independently of each other, assumed in Chapter 5, and under the assumption that the population is sufficiently large, according to Vanberkel (2011). The probability mass function of a Poisson distribution is defined as:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Here, $k \geq 0$ is an integer and the parameter λ is the expected value of the variable X.

This distribution is chosen since it enables the simulation model to randomly draw an integer number of arriving patients in a time period only based on the expected value. In the simulation model, as in the optimization model, a time period equals one day. Thus, the Poisson distribution draws the number of arrivals that occurs within a day. The expected number of arrivals is based on the same assumptions as introduced in Section 8.1.1, under the heading "Patient demand". In Table 11, the λ-values for the relevant cancer types are shown.

Table 11: λ-values for the Poisson distribution.

Care pathway	Value of λ
Uterine	0.8
Cervical	0.6
Ovarian	0.8

Figure 18 illustrates the shape of the probability mass functions of the Poisson distribution with the λ values from Table 11.

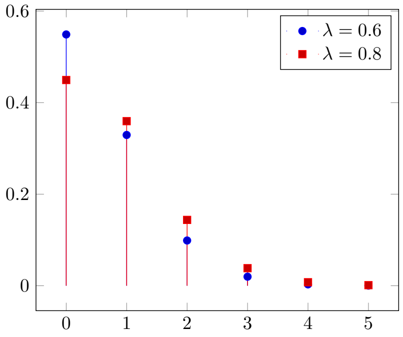


Figure 18: Poisson distribution of patient arrivals.

Service time

The service time of an activity is drawn from a discrete probability distribution. The distribution is designed based on observations of service times at the Oslo University Hospital, and it is found that the service times are dependent on the associated queue j and tend to spread symmetrically around a mean value μ_j. One could also choose a normal distribution due to its symmetry, but a discrete distribution is rather chosen in order to limit the number of possible outcomes and due to the fact that probabilities are non-negative numbers.

Table 12: Probability distribution of service times.

Service time	Probability
$0.6\mu_j$	0.1
$0.8\mu_j$	0.2
μ_j	0.4
$1.2\mu_j$	0.2
$1.4\mu_j$	0.1

It is noted that the expected value of the entries in Table 12 is μ_j:

$$\mathbf{E}(\text{service time}) = 0.1 \cdot 0.6\mu_j + 0.2 \cdot 0.8\mu_j + 0.4 \cdot \mu_j + 0.2 \cdot 1.2\mu_j + 0.1 \cdot 1.4\mu_j = \mu_j$$

Transition from diagnosis to treatment

The transition from the diagnosis stage to the treatment stage poses the only uncertain split in a patient's care process. After the diagnosis, a patient may either enter her treatment phase, or be discharged. The latter is the case if no cancer is detected in the diagnosis phase. When a patient leaves the diagnosis stage and enters the treatment stage, we do not have any information available for this patient in specific on which treatment this patient shall receive. Therefore, our best estimate is to assign treatment to patients based on a discrete probability distribution based on historical data. This distribution is based on the probabilities that are expressed in the parameters P_{ij}, described in Section 8.1.1 .

In the second column of Table 13, the probabilities of entering the different possible treatment stages are given. Note that there are three different treatment processes a patient with cervical cancer may be assigned to, and that the corresponding number for uterine and ovarian cancer is two. The probability of discharge, found in the third column, also varies for the three. It is noted that the row-wise sum adds up to 1.

Table 13: Probability distribution of treatment stages.

Care process	Probability of path	Probability of discharge
Uterine	0.15, 0.35	0.50
Cervical	0.27, 0.03, 0.60	0.10
Ovarian	0.25, 0.65	0.10

No show

Whether or not a patient shows up to an appointment in order to be serviced, is a question that has only two possible outcomes. The Bernoulli probability distribution, which is equivalent to a binomial probability distribution with $n = 1$, is suitable and it is defined as follows:

$$P(X = k) = p^k (1 - p)^{1-k} \qquad\qquad k = 0, 1$$

for $0 < p < 1$ where p is the probability of a no-show. $k = 1$ corresponds to a no-show and $k = 0$ corresponds to a patient showing up. We have no information from Oslo University Hospital on the distribution of no-shows, but we assume that the value of p takes the values stated in Table 14. However, simulations using different values of p are assessed in Section 9.2.

Table 14: Probability distribution of no-show.

No-show	Probability
p	0.05
$1 - p$	0.95

8.2.2 Patient scheduling policy

There are different ways of choosing which patients from a queue to service first. From a medical perspective, there will always be some patients that require more urgent care than others. However, it is hard to measure the degree of urgency exactly, and some guiding principles for the scheduling policy must be made.

We assume a scheduling policy where the patients that have waited the longest in the system in total, are selected first, as introduced in Chapter 6. A strength of this policy is that it is considered somewhat fair, i.e. the patient that has waited the longest, is serviced first. However, this method lacks the ability to prioritize patients that require urgent care from a pure medical perspective.

Alternatives to this policy, is the FIFO policy, where the patient that has waited the longest in each queue is serviced first. Another alternative is to always service the patient in a queue that is closest to a time limit in her care process.

8.2.3 Warm-up period

The simulation is initialized from an empty system, and it takes some time periods of simulation before the system reaches its steady state. The time periods from the simulation starts and until this steady state is reached, is called the warm-up period of the system. The exact length of this warm-up period is dependent on the size of the instance that is being simulated. By visualizing the total number of patients queuing in the system, as done in Figure 24, it becomes evident that the warm-up period lasts until the graph flattens out after a period of rapid increase up until this. As a general rule of thumb for the cases discussed in Chapter 9, we assume 7 weeks of simulation to be a sufficiently long warm-up period.

9 Computational study

In this chapter, results from a computational study of the optimization model presented in Section 5 and the scheduling framework presented in Chapter 7 are presented and discussed. It is underlined that the scheduling framework includes both the optimization model and simulation model, as visualized in Figure 16.

A technical study is performed in Section 9.1. Next, managerial insights from running the scheduling framework by altering parameters and simulation characteristics are described in Section 9.2. Finally, a case study of the Department of Gynecological Cancer at Oslo University Hospital is performed in Section 9.3.

It is important to notice that the hospital is not going to conduct the simulations presented in Sections 9.2 and 9.3. The simulations are only used in this thesis to evaluate the schedules generated by the optimization model. Thus, the scheduling framework acts as an imitator of how the hospital may put the optimization model to use.

The optimization and simulation models are both implemented in the programming language Python. For the optimization, an interface to the commercial optimization software Gurobi 8.1.1 was used. The instances were solved on a computer with Intel Core i7-7700 (3.60 GHz) CPU and 32 GB RAM, running on the Windows 10 64-bit operating system.

9.1 Technical study

In this section, we perform a technical study of the optimization model. Here, no simulations are involved, and the objective is solely to test the optimization model in isolation. To perform the technical study, we employ the input data presented in Section 8.1. The main purpose of this section is to investigate how the optimization model responds when it is run using different instances and when altering key input parameters. In Section 9.1.1, the instances used in the technical study are presented. Thereafter, Section 9.1.2 presents the results when the instances are run using the optimization model. Finally, in Section 9.1.3, two patient admission schedules generated by the optimization model are presented.

9.1.1 Instances used in the technical study

Each instance in Table 15 is described by an instance name, its associated cancer care pathways, number of queues and the length of the planning period. The characters in the instance name describe which cancer care pathways the instance consists of, using the first letter of each pathway as an identification. The number at the end of the instance name represents the length of the planning period. Thus, UCO-14 means that the instance consists of the uterine, cervical and ovarian care pathway, and the planning period covers 14 time periods.

What separates the instances, is the number of queues and the length of the planning period. These two sets are chosen because it is the number of time periods and queues that affects the complexity of the model the most. The number of queues is determined by the number of care pathways and the number of activities associated with these care pathways. The number and sequence of the activities within a care pathway is not being altered, because changing the number of activities within one care pathway would result in a more vague and less intuitive problem instance. Instead, changing the number of cancer care pathways is a more clear and unambiguous way to increase the total number of queues in the system. Thus, increasing the number of care pathways implies increasing the number of queues. In addition, since there is a demand associated with every care pathway, the total of patients in the system increases when the number of different cancer care pathways included in the system is increased.

Table 15: Overview of instances tested in the technical study.

Instance	Care pathway	Queues	Time periods
U-7	Uterine	11	7
U-14	Uterine	11	14
U-21	Uterine	11	21
U-28	Uterine	11	28
U-35	Uterine	11	35
U-42	Uterine	11	42
U-49	Uterine	11	49
UC-7	Uterine, cervical	28	7
UC-14	Uterine, cervical	28	14
UC-21	Uterine, cervical	28	21
UC-28	Uterine, cervical	28	28
UC-35	Uterine, cervical	28	35
UC-42	Uterine, cervical	28	42
UC-49	Uterine, cervical	28	49
UCO-7	Uterine, cervical, ovarian	40	7
UCO-14	Uterine, cervical, ovarian	40	14
UCO-21	Uterine, cervical, ovarian	40	21
UCO-28	Uterine, cervical, ovarian	40	28
UCO-35	Uterine, cervical, ovarian	40	35
UCO-42	Uterine, cervical, ovarian	40	42
UCO-49	Uterine, cervical, ovarian	40	49

9.1.2 Results from the technical study

As mentioned in Chapter 5, the problem to be solved is modeled as a Mixed Integer Programming (MIP) model. The Gurobi solver uses the Branch and Bound (B&B) algorithm for solving such problems. Due to the complexity of MIPs, the solution time may be long despite of convenient instance sizes. In order to cope with this challenge, the algorithm is set to stop at a certain time, since running the model until an optimal solution is found is not always necessary, nor possible. In this technical study, the instances are run with a maximal time limit of 10,800 seconds. This time limit was chosen based on the fact that instances reaching this limit seldom improved much if they were run for a greater amount of time. Also, if the model was to be implemented for use in a real-life setting, three hours of waiting is already too long for all practical purposes. If an instance is not solved to optimality within the time limit, the time of the associated run is marked with an asterisk (*). Furthermore, if a problem instance reaches the time limit when it is executed, the objective value of the best feasible solution satisfying the integer constraints, is enlisted in the objective value column in Table 16. The LP-gap, the difference in percentage between the best feasible solution found and the LP-relaxation of the problem, is also recorded if the model is stopped before an optimal solution is found. The instances where no optimal solution is found are included in the technical study because they may also provide valuable insights on how the model functions and define an upper bound on how large instances we are able to solve to optimality.

In Table 16, an overview of the results when running the instances described in Table 15 is shown. In every run, the objective value, solution time, LP-gap and the number of constraints, continuous and integer variables and B&B-nodes are recorded. Each run of the instances is starting out from an empty system. This makes the problem somewhat faster to solve compared to if the system was initiated with patients already residing in the queues. This is because there are more patients to service and therefore a greater allocation problem. Therefore, one should anticipate the solution times to be some longer when the model is run as a part of the scheduling framework, but the differences are not big enough to make a great impact.

Table 16: Technical summary of the optimization model. When performing the technical study, we run each test instance five times, and the values recorded are mean values from these runs.

Instance	**Obj. value**	**LP-gap [%]**	**Time [s]**	**Con- straints**	**Cont. variables**	**Int. variables**	**B&B- nodes**
U-7	22.3	0.0	0.4	158	247	34	1
U-14	76.1	0.0	1.1	655	1,972	89	164
U-21	129.0	0.0	5.1	1,497	6,388	141	2,057
U-28	194.0	0.0	38.7	2,758	14,980	196	6,359
U-35	255.8	0.0	116.4	4,440	28,942	251	10,550
U-42	282.6	0.0	694.1	6,545	49,399	306	53,343
U-49	336.4	3.0	10,800*	9,041	76,464	361	366,569
UC-7	37.7	0.0	1.0	234	378	54	1
UC-14	145.1	0.0	3.0	1,109	3,266	156	548
UC-21	280.0	0.0	15.0	2,719	11,281	275	3,188
UC-28	451.4	0.0	114.0	5,336	27,894	410	41,578
UC-35	627,8	0,0	1450.8	8,969	58,400	570	115,766
UC-42	771.9	2.6	10,800*	13,641	103,147	714	277,217
UC-49	1,248.2	47.7	10,800*	22,102	161,604	830	693,080
UCO-7	59.6	0.0	1.4	349	561	79	14
UCO-14	206.9	0.0	7.4	1,613	4,758	230	2,206
UCO-21	387.9	0.0	70.4	3,948	16,493	409	12,932
UCO-28	608.0	0.0	1828.0	7,672	40,683	604	220,404
UCO-35	833.5	6.2	10,800*	12,967	82,052	804	360,292
UCO-42	1,280.3	35.8	10,800*	22,984	147,957	1,004	508,304
UCO-49	2,010.0	54.8	10,800*	32,257	234,987	1,205	784,369

The results in Table 16 show that the objective value and solution time increase when the number of care pathways is increased. To compare the different set of pathways in a fair manner, the number of time periods must be isolated in order to register changes. Thus, if the instances U-7, UC-7 and UCO-7 are compared, the results show that both the objective value and solution time increase with the number of queues. The objective values for the given instances are 22.3, 37.7 and 59.6, respectively. Thus resulting in an increase of about 60 % to 70 % for every care pathway that is added. The objective value increases with the number of care pathways, because there are more queues and also more patients demanding services. These factors increase the probability of patients residing in queues for more time periods, thus increasing the objective value, which is a weighted sum of total waiting time in the system. Additionally, the solution time increases from 0.4 for the U-7 instance to 1.0 for the UC-7 instance and 1.4 for the UCO-7 instance. The reason for the increased solution time comes from the greater problem complexity resulting from the increased number of care pathways.

The objective value and solution time also proves to increase with the length of the planning period. In order to measure the change in the objective value and solution time, a set of care pathways is isolated, in this case we look into the instances UCO-7, UCO-14, UCO-21, UCO-28, UCO-35, UCO-42 and UCO-49.

The results show that the objective value increases gradually from 59.6 for the UCO-7 instance to 2,010.0 for the UCO-49 instance when the length of the planning period is increased. There are more patients that enter the system when the number of time periods is increased, thus the overall number of patients that resides in the system is increased and thus the number of time periods patients are waiting in queue may increase. Therefore, the objective value rises with the number of time periods. Additionally, the solution time increases exponentially from 1.4 seconds for the UCO-7 instance to the time limit of 10,800 seconds for the UCO-35, UCO-42 and UCO-49 instances. Thus, a longer planning period increases the complexity of the model to a great extent.

The number of constraints and continuous and integer variables presented in Table 16 are from after a presolve of the mathematical model is done. The presolve is a method that transforms the original model into a smaller, equivalent model. Hence, the problem becomes easier to solve. Since it is the model after the presolve that is actually optimized, enlisting the number of constraints and integer and continuous variables after this operation, instead of before, gives a more realistic view of the problem complexity. We observe that the number of constraints and continuous and integer variables increases with the number of care pathways and planning periods, which underpins what is discussed above.

Finally, the number of B&B-nodes indicates the number of nodes in the B&B-tree that is explored before an optimal solution is found, or the time limit is reached. Thus, the number of B&B-nodes does not represent the number of nodes that were needed to find the optimal solution. Therefore, as with the solution time, the number of B&B-nodes also serves as a measure of the complexity of the problem instance. In accordance with the previous results, the number of B&B-nodes increases with the number of time periods and care pathways thus implying an increased problem complexity.

9.1.3 The patient admission schedules from the optimization model

As covered in Chapter 5, the mathematical model returns schedules that route patients through their associated care processes. The schedules are derived from the variables b_{jt}, which denote the number of services to be conducted for a given queue j at time period t. The Figures 19a and 19b, are snapshots of two such schedules generated by the mathematical model solved for the U-14 instance (see Table 15) in a simulation. The schedules are extracted from the simulation and not from a single run of the optimization model, because the schedules from the simulations enable us to see the connections between two consecutive schedules. The implementation period of the instance spans 7 time periods and the planning period consists of 14 time periods. The implementation period defines the number of time periods that are to be implemented between every run of the optimization model.

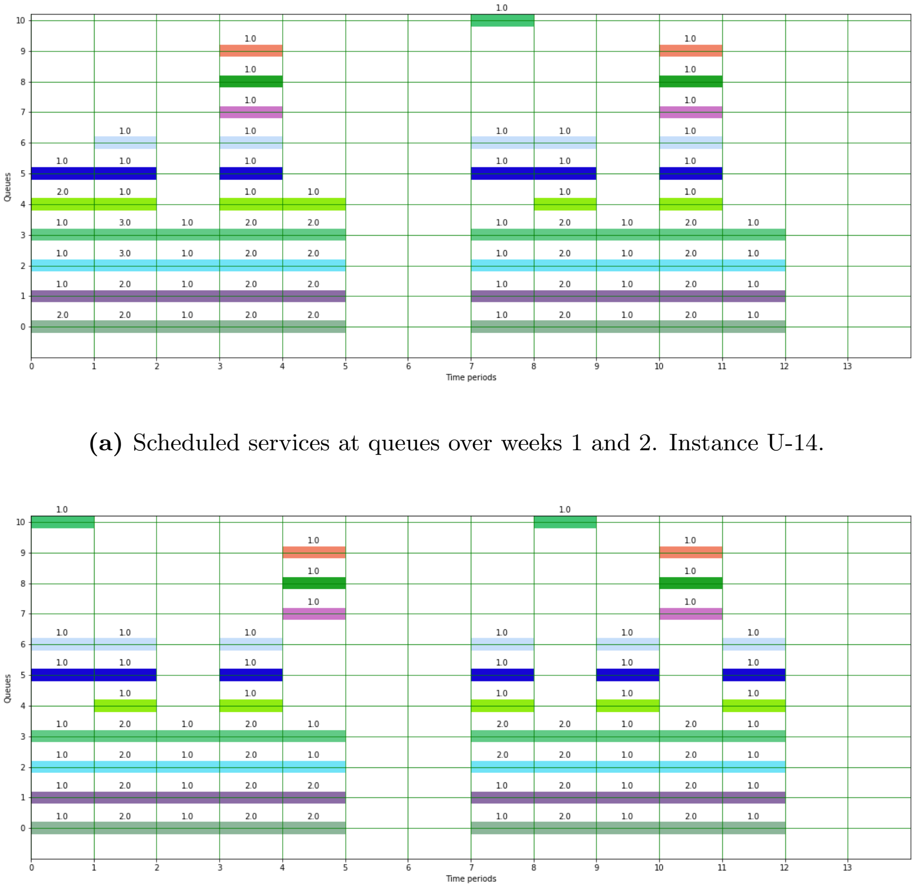
Each row in the two figures corresponds to a queue, while the columns define the time periods. If there is scheduled a service at a queue in a given time period, the associated time period and queue is colored and numbered. The color is unique for every queue, and the number denotes the number of patients that are scheduled to be serviced from the associated queue in the given time period. The queues 0 to 3 belong to the diagnosis stage of the uterine cancer care pathway, whereas the queues 4 to 6 and 7 to 10, belong to the first and second treatment paths, respectively. Note that no services are conducted in the time periods 5, 6, 12 and 13, since it is assumed that there are no services during weekends.

The mathematical model is solved using a rolling horizon approach, where the first schedule makes a plan for week 1 and 2, seen in Figure 19a. Since there is an implementation period of 7 time periods, only the first week is therefore implemented. After the week is implemented, the optimization model is run once more in order to create a plan for week 2 and 3, seen in Figure 19b. Since there already exists a plan for week 2 from the first schedule, the new schedule accounts for this by using the previous generated schedule for week 2 as a basis, but with some allowed alterations. If we look at the second week in the first schedule in Figure 19a, and compare it to the first week in the second schedule in Figure 19b, one may observe the aforementioned dynamics. The second schedule makes some adjustments to the previous generated schedule for week 2. For example, queues 0 and 1 are servicing one more patient each at time period 4 in the updated schedule. Furthermore, the appointments for queues 7 to 9 in the fourth time period in the second week in the first schedule have been transferred to the next time period in the new schedule.

From the figures we can observe that the optimization model, at this specific point of time in the simulation, schedules a lot of services to the queues in the diagnosis phase. As one may see, in time periods 0 to 5 and 7 to 11, all the diagnosis queues have scheduled services in both of the schedules. This is due to the fact that the diagnosis queues have more patients entering the queues, since the number of patients in the treatment paths is based on a fraction of the patients from the diagnosis phase. Furthermore, we also see that the first treatment path, queues 4 to 6, is assigned more scheduled services compared to the second treatment path, queues 7 to 10. This difference is due to the fact that the two treatment paths have a different shares of patients entering the respective paths after the diagnosis phase, decided by the parameters P_{ij}. For the first treatment path, the corresponding parameter, P_{ij}, has a value of 0.35, while as for the second treatment path the value is 0.15, thus explaining why there are more scheduled appointments for the first path. The remaining 50 % of the uterus cancer patients are discharged after diagnosis stage with no cancer.

It is also possible to see in Figure 19, how the model efficiently routes patients when there are no delays between the activities. For example, in Figure 19b, the queues 4 to 6 in the time periods 7, 9 and 11, have scheduled three services right after one another in the same time period. This is also a recurring pattern in other time periods. The activity associated with queue 4 is a visit at the outpatient clinic, whereas the activity associated with queue 5 is a gynecological examination. These activities are of short duration and do not result in a delay after the activity is undergone, unlike surgery. Thus, the model schedules these activities with low spacing in order to get the patients efficiently through their associated pathway in an efficient manner, which reduces the overall waiting time.

These sequences of activities which take place the same day, may pose some challenges when the schedules are tested and simulated in the scheduling framework. A sequence of activities during a single time period causes strong dependencies, and when uncertainties are introduced, for example no-shows, the sequence is easily broken and resource capacities are lost.



(a) Scheduled services at queues over weeks 1 and 2. Instance U-14.

(b) Scheduled services at queues over weeks 2 and 3. Instance U-14.

Figure 19: Scheduled services at queues over three weeks. Instance U-14.

9.2 Managerial insights

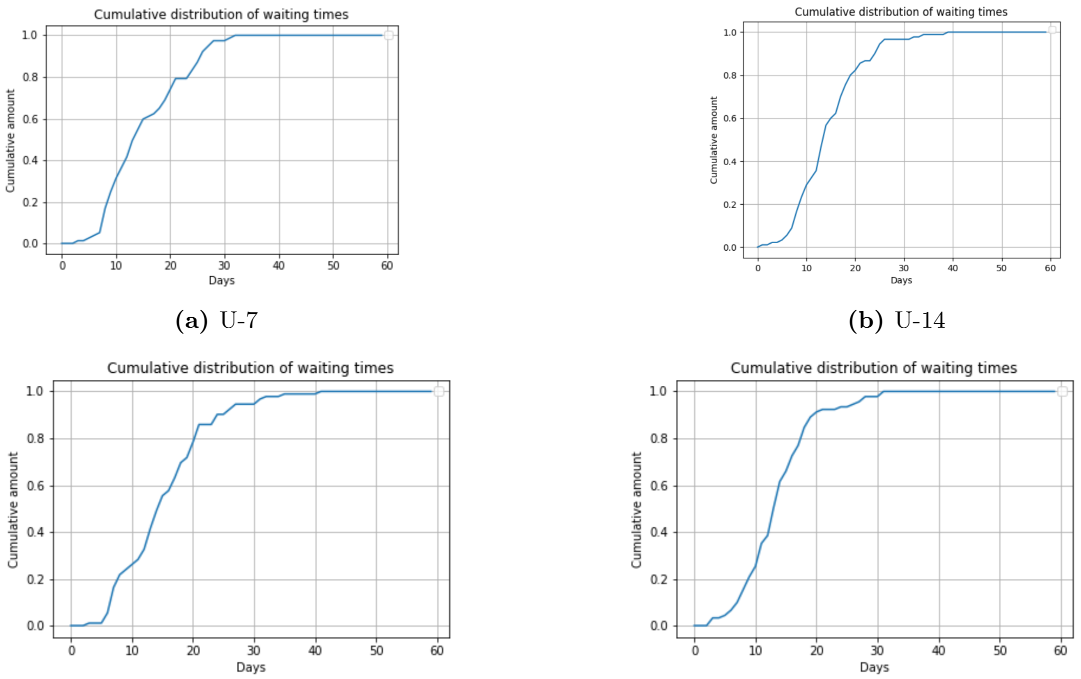
In this section, we present managerial insights from running the scheduling framework presented in Chapter 7. The insights provided in this section is of interest to hospital planners and management, as they reveal how our work may help making well-informed decisions. In order to observe the dynamics of the system over time, we choose a simulation period of 45 weeks. The simulation model is run for 7 weeks prior to the simulation period as a warm-up period until a steady state is reached. Any data generated during these 7 weeks are neglected in the analyses in this chapter. However, in figures that visualize how the system develops over the simulation period, the warm-up period is included for visualization purposes only. The implementation period is 7 days, if not stated otherwise, thus the optimization model is run once every week.

In Section 9.2.1, we assess how the scheduling framework reacts when the length of the planning period is altered. Secondly, in Section 9.2.2, we assess how the scheduling framework reacts when the length of the implementation period is changed. Section 9.2.3 evaluates how the number of allowed shifts in the rolling horizon approach affects the system performance. Finally, in Section 9.2.4, the effect the no-show rate has on patient waiting time is evaluated.

9.2.1 Varying the length of the planning period

In this section the effects of varying the length of the planning period is discussed. The planning period defines the set of time periods the optimization model solves the problem for. When testing different planning periods, we use the instances U-7, U-14, U-21 and U-28. These instances are chosen because they are of low complexity and thus providing reasonable solution times.

The average waiting times for the different instances is about 15 days, but Figure 20 shows to a greater extent how these waiting times are distributed. The figure shows the cumulative distribution of waiting times for the instances U-7, U-14, U-21 and U-28. As one may see, the distributions are mostly similar, but as the length of the planning period increases, the graphs tend to delinearize and bend where the waiting period is equal to 20. For example, 80% of all patients wait for 21 days or less in the U-7 instance. However, when the planning period is increased to 28 days, 80 % of all patients wait in total 17 days or less. This means that the waiting time for the majority of the patients decrease when the planning period is increased. The average waiting times are about the same because the instance with longer planning periods also tend to have the patients with the longest waiting times compared to the instances with lower planning periods.



(a) U-7

(b) U-14

(c) U-21

(d) U-28

Figure 20: Cumulative distribution of waiting times when varying the length of the planning period.

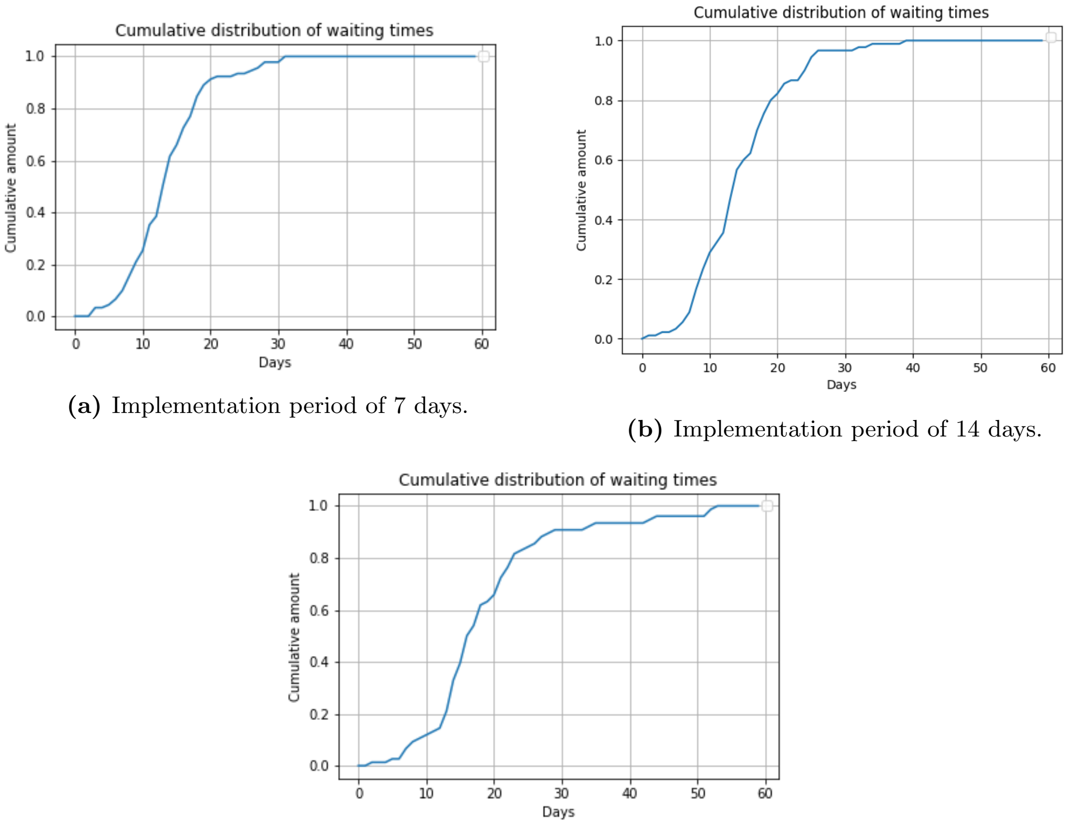
The model prioritizes to service patients who have been in the system the longest, and when the planning period is increased, the model looks further ahead in time, identifying earlier the patients that are going to accumulate many waiting periods. This way a longer planning period entails more proactive schedules. However, schedules with a long planning period also result in some patients to wait longer. A reason for this is that planning longer ahead in time accounts less for the day-to-day fluctuations, thus the schedules are less flexible and it may lead to a few patients waiting longer.

Furthermore, as pointed out in Section 8.1, another positive outcome of an increased planning period is more predictability for the staff and patients at Oslo University Hospital. However, because changes of appointments scheduled after the implementation period are allowed, the predictability depends on how the parameter defining the number of shifts is set. If the number of allowed shifts is low, more predictability is achieved. Thus, a long planning period is not tantamount to high predictability. In order to achieve the most predictability, one should instead increase the length of the implementation period as this freezes the schedules without any option to alter them afterwards. The effects of varying the implementation period are discussed in Section 9.2.2.

9.2.2 Varying the length of the implementation period

In this section, the effects of changing the length of the implementation period is discussed. For this purpose, we use the U-28 instance, due to its convenient solution time. The instance is run with implementation periods of lengths 7, 14 and 21 days.

The cumulative waiting time distributions for the three different lengths of implementation periods are shown in Figure 21. When the length of the implementation period is increased, the optimization model is run a less number of times, because the model is only run before each implementation period. Therefore new plans are less frequently generated, the model adjusts less frequently to the system state and it is less flexible. As seen in Figure 21, the distributions shifts more to the right as the implementation period is increased and the model gets less flexible. In Figure 21a, where the implementation period is 7 days, all patients wait for 30 days or less. On the other hand, when the implementation period is increased to 21 days in Figure 21c, we see that there are patients with waiting times of over 50 days. Thus, when the implementation period increases, the model becomes less flexible, resulting in longer waiting times. On the other hand, a longer implementation period also gives some advantages for the people involved. A longer implementation period yields more predictability, as the schedules are frozen for a longer time.



(a) Implementation period of 7 days.

(b) Implementation period of 14 days.

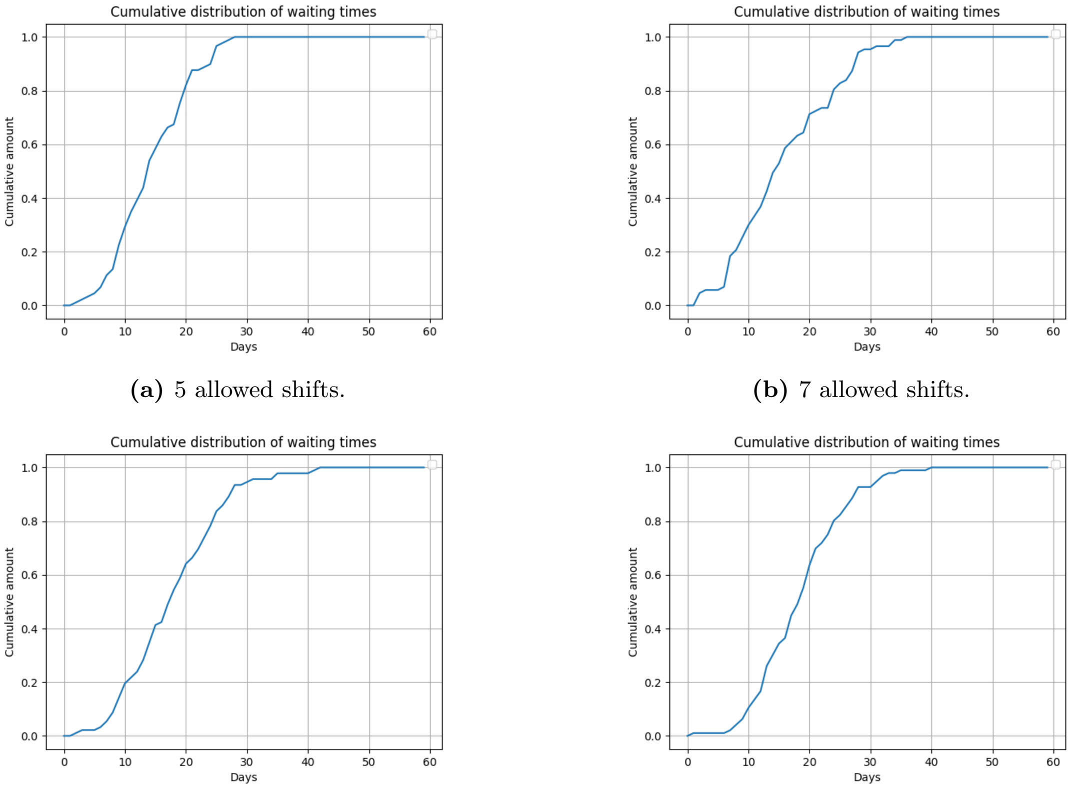
(c) Implementation period of 21 days.

Figure 21: Cumulative distribution of waiting times when varying the length of the implementation period.

9.2.3 Varying the number of allowed shifts in rolling horizon

In the following, the effect of varying the number of allowed shifts is examined. The number of shifts determines the allowed number of changes in the new appointments generated by the optimization model, compared to the previous schedule for time periods that are overlapping. A low value implies that the newly generated schedules need to follow the previous plan in a strict way, while a high value means that the new plan may pay less attention to the previous one. In the following, we use the U-14 instance, with an allowed number of daily shifts of value 5, 7, and 9. For comparison reasons, an example where the number of allowed shifts is fully relaxed is also included.

In Figure 22, we see that the waiting times shift to the right as the number of allowed shifts increases. Furthermore, the average waiting times for the four different cases are 14.9, 16.3, 18.1, 19.1, respectively. These results show that when the number of shifts increases, the waiting time for the patients increases. A low number of shifts implies a more restrictive plan which forces the model to a greater extent adhere to the previous generated plan. In this way, the model is less myopic and less affected by small changes and variations. Therefore, as the number of allowed shifts increases, the model adheres less to the previously generated plans, which causes the model to be more affected by short term fluctuations, and as it shows in Figure 22, this leads to increased waiting times.



(a) 5 allowed shifts.

(b) 7 allowed shifts.

(c) 9 allowed shifts.

(d) Unlimited amount of allowed shifts.

Figure 22: Cumulative distribution of waiting times when varying the number of allowed shifts.

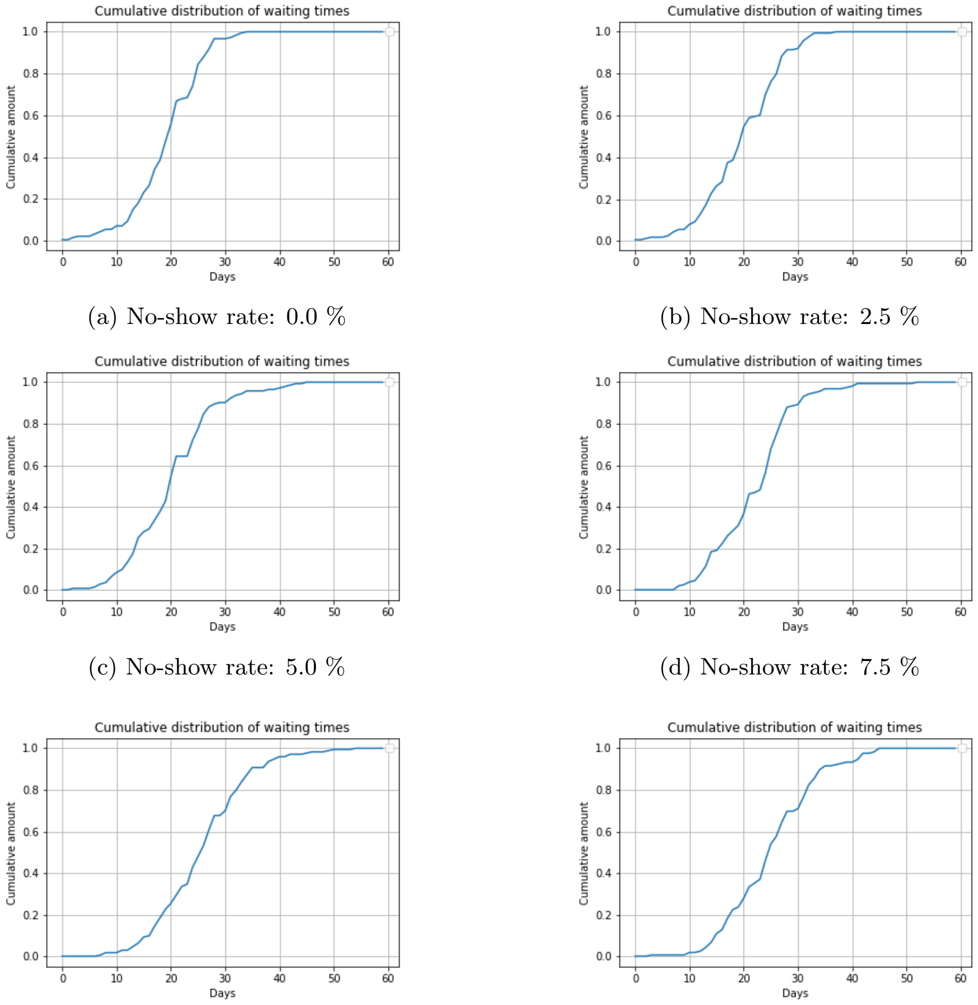
9.2.4 Evaluating waiting time with different rates of no-show

In Figure 23, the cumulative distribution of waiting times for the U-14 instance with different no-show rates are illustrated. This instance is significantly smaller than the instance UCO-14 that replicates the situation at the Department of Gynecological Cancer. A summary of the cumulated amount for each no-show rate is provided in Table 17. The motivation behind testing different rates of no-show on a small test instance, is to conceptualize what happens to the waiting time when no-shows increases, and evaluate how the system performs when the no-show rate increases.

We see that the graph moves slowly from left to right when the no-show rate increases. With a no-show rate of 0 %, all patients in the simulation have finished their entire care process within 33 waiting days, whereas the equivalent number for a no-show rate of 10 % is 53 days.

The graph for lower no-show rates are steeper than of those for higher no-show rates. With low rates of no-show, there is little need to re-schedule patients. Contrarily, when the no-show rate increases, more patients needs a re-scheduling and waiting times increases. The more the no-show rate increases, the more apparent is the increase in waiting times.

We observe that the waiting time needed to accumulate 100 % of the patients for a no-show rate of 12.5 % is 44 days. This number stands out from the rest of the column it belongs to, where the numbers increase, more or less, steadily with increased rates of no-show. We are not able to explain how this result came into being. However, similar results appeared over and over again when the simulation was done repeatedly.



(a) No-show rate: 0.0 %

(b) No-show rate: 2.5 %

(c) No-show rate: 5.0 %

(d) No-show rate: 7.5 %

(e) No-show rate: 10.0 %

(e) No-show rate: 12.5 %

Figure 23: Cumulative distribution of waiting times with different rates of no-show for U-14 over 45 weeks.

Table 17: Waiting time for combinations of cumulated amount and no-show rates from Figure 23.

		Cumulated amount					
		0 %	20 %	40 %	60 %	80 %	100 %
No-show rate	0.0 %	0	14	18	20	24	33
	2.5 %	0	13	17	22	26	36
	5.0 %	0	13	18	21	26	43
	7.5 %	0	15	20	24	26	52
	10.0 %	0	18	24	27	32	53
	12.5 %	0	17	23	26	32	44

9.3 Case study

This thesis seeks to solve the problem of scheduling patients in standardized cancer care processes at the Department of Gynecological Cancer at Oslo University Hospital using operations research methods. In order to evaluate how the system performs in a more realistic setting, we provide in the following section a case study inspired by the hospital. The test instance used is UCO-14, i.e. a case with uterine, cervical and ovarian cancer patients where the planning period is 14 days and the implementation period is one week. It is underlined that a length of the implementation period of one week is somewhat short if the model was to be implemented in a real-life setting. The employees need to be able to plan their working life more than a week ahead. The short implementation period was chosen based on computational complexities when increasing its length - the runtime for the optimization model is 10 times higher for UCO-21 compared to UCO-14. In the simulations conducted in this section, we simulate over a period of 45 weeks. The warm-up period is set to 7 weeks.

In this section, we assess how the simulation model schedules patients, how many patients that are in the system in the simulation period and the patient waiting times. The input data to the optimization model, is found in Section 8.1. Input data for the simulation model, is found in Section 8.2. It is underlined that the input data to the case study is subject to a high degree of uncertainty. As previously stated, we lack proper input data to some key parameters. Consequently, the case study presented in this section is inspired by Oslo University Hospital and not to be seen as a sterling implementation. Due to lack of data, we are unfortunately not able to compare our model to the real-life situation at the hospital.

In Section 9.3.1, the, main results from running the scheduling framework for the case from Oslo University Hospital are stated. An assessment of the number of patients that resides to the system during the simulation period is given, and we assess the waiting times for the patients in the system. Section 9.3.2 evaluates the effects of changing the scheduling policy. Finally, reflections on the results are presented in Section 9.3.3.

9.3.1 Evaluating the optimization model using the scheduling framework

In this section the results from from running the scheduling framework on the case data from Oslo University Hospital are presented. As explained earlier, the optimization model generates schedules with plans on how many patients that are to be serviced at the different queues for all days in the planning periods. For this case, the number of different queues, is 40. These schedules are then realized by a simulation model. The interplay between the optimization and simulation model is encapsulated in the scheduling framework presented in Chapter 7. In Table 18 below, key values from the case study are presented.

Table 18: Key values from the case study.

Description	**Value**
Number of patients generated	553
Number of patients that have undergone treatment	365
Number of patients discharged right after the diagnosis stage	131
Average patient waiting time	34 days

Development of total patients in the system

This section looks into how the number of patients in the system develops throughout the days that are simulated in the scheduling framework. Over the simulation period, 553 patients enter the system. Of them, 365 patients have received a treatment and 131 patients did not have cancer and were discharged right after the diagnosis stage.

In Figure 24 the development of the number of patients in the system is graphed. The figure displays two lines, one indicating the actual number of patients residing in the system at a given day and one representing the moving average. This moving average is calculated by taking the average of the previous n data. A moving average is used because it smooths out short-term fluctuations and highlights trends. Here, the value of n is set to 14, i.e. it shows the mean number of patients in the system for the last two weeks.

For the line representing the number of patients in a queue, we identify small plateaus appearing with a fixed frequency of one week. This indicates some sort of seasonality. This is caused by the fact that no patients leave or arrive to the system during weekends. Thus, the number of patients remains the same over the weekends.

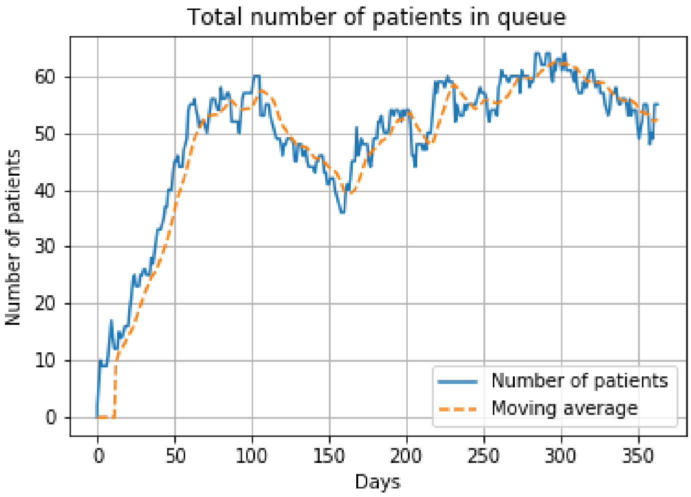


Figure 24: Number of patients in the system. The warm-up period makes up the first 7 weeks.

In Figure 25, the development of the number of patients in the system are shown for each type of cancer. Throughout the simulation period, the uterine pathway has in average the lowest number of patients residing in the pathway. There may be several causes for this. One of the reasons may be that the uterine pathway has 11 queues, which makes it the pathway with the least amount. Another reason is that it also has the highest discharge percentage of all of the pathways where about 50 % of the patients are discharged after the diagnosis stage. Furthermore, the activities in its two associated treatment pathways do not require long delays of medical reasoning. Therefore, the number of patients associated with this care pathway is lower compared to the other care pathways.

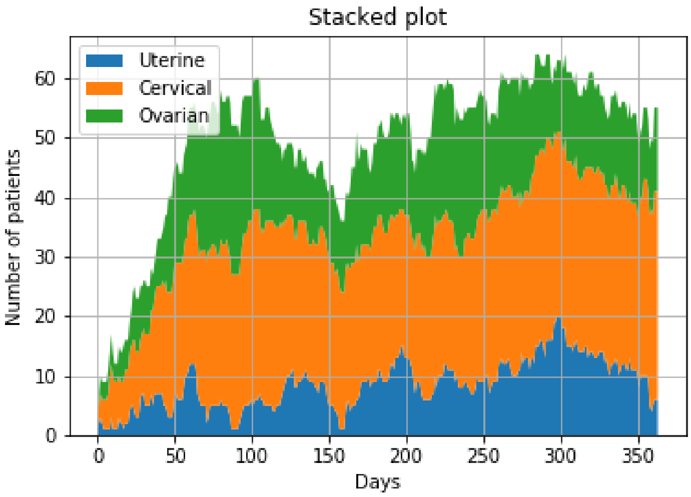


Figure 25: Number of patients in the system per patient group.

Figure 25 also shows that the cervical cancer care pathway mostly holds the highest number of patients. The reasons for this are much the opposite of those mentioned in the case of the uterine care pathway. Firstly, this pathway includes 17 queues, which makes it the one with the most queues. This leads patients to stay longer in the system because they have to undergo more activities. Also, the activities in the different forms of treatment associated with this cancer pathway are demanding relatively long medical reasoned delays after they are undergone. For instance, surgery and several rounds of chemotherapy are activities causing greater delays. Since many patients reside in this care pathway, it may also affect the waiting times for the patients with this diagnosis.

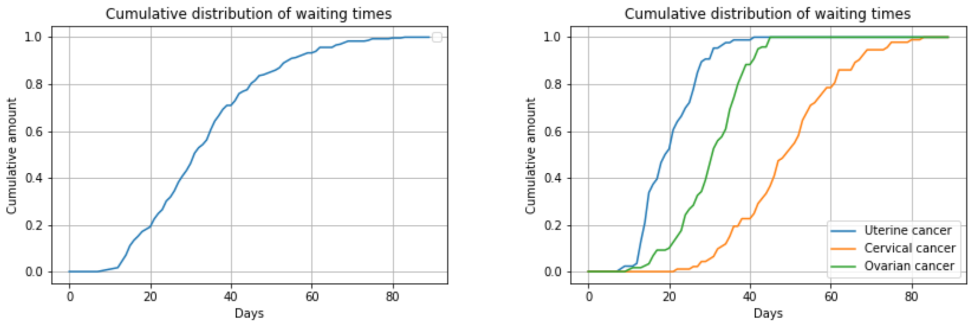
The number of patients residing to the ovarian cancer care pathway fluctuates somewhere between the two other pathways. There are 12 queues associated to this care pathway, therefore it does not differ much from the number of queues in the uterine pathway. One of the reasons why there is on average a higher number of patients associated to this care pathway, is due to the fact that the delay times after activities in the treatment stage are relatively higher compared to those in the uterine care pathway. This is because the ovarian cancer is considered the most lethal cancer disease compared to the other two gynecological cancers, requiring thorough procedures and more demanding activities which leads to patients residing at the associated queues for a longer time.

It is underlined that a care process that consists of relatively many activities, should not automatically have more patients residing to its queues than others. To motivate a behavior where the number of patients queuing for the activities in the different care processes are smoothed out, one may adjust the objective function weights. This adjustment may either be done manually, or by using the alternative procedure proposed in Appendix B.

Waiting times

The waiting time for patients in the system is affected by the number of patients already in the system when the patients arrive. In this thesis, the waiting time is defined as the time a patient spends in a queue, i.e. time between activities. It is noted that the waiting times are based on the patients that have completed both the diagnosis and treatment stages, and not patients that are discharged after the diagnosis stage. The ovarian and uterine cancer patients have the lowest average waiting times, with values of 20 and 31 days, respectively. For these two cancer types, queues do not build up in the same way as for cervical cancer (Figure 25), and we are able to keep the waiting time low. Furthermore, the cervical cancer patients have to wait the longest, with an average waiting time of 50 days. Additionally, the overall average waiting time for all of the patients is 34 days.

In Figure 26a we observe the cumulative distribution of waiting times for all of the patients that have exited treatment. From the figure, it becomes evident that 50 % of the patients experience waiting times shorter than 30 days and the patient population as a whole has waiting times shorter than 82 days. Figure 26b shows how the cumulative distribution of waiting times for each of the care pathways, and it shows in more detail the underlying contributions to the cumulative distribution in Figure 26a. The steep, almost parallel green and blue lines explain why the uterine and ovarian care pathways have the lowest waiting times, where in practice, the majority of the ovarian and uterine cancer patients experience less than 40 days of waiting. The orange line, representing the cervical cancer patients, is more flattened. This implies that the distribution of waiting times is more evenly spread, thus entailing a higher variance of waiting times. This may come from the fact that the cervical cancer patients may follow three different treatment paths, where each of the treatment paths has lengths of 1, 3 and 8 queues. Because of the different lengths of the treatment paths and that the waiting times in general increase with the length of the care pathway, the waiting times for the cervical cancer pathways are distributed in a more wide range, giving the less steep, orange line.



(a) All patients aggregated.

(b) Each cancer type separated.

Figure 26: Cumulative distribution of waiting times.

In addition to generation of patient admission schedules, the optimization model also allocates resources accordingly. In the following, the thesis looks into how the resources, activities and waiting times are interconnected.

In order to assess this interconnection, we need to know how the activities and resources are distributed over the different pathways. Table 19 enlists the number of occurrences of the different activities for each of the care pathways. The number of occurrences is the aggregated sum of all the activities in all of the paths within a cancer care pathway in order to give an idea on how the activities are distributed. Since the activities are associated with the same set of resources for the different care pathways, it gives an indication of the degree to which it is allocated. On the other hand, it is not a realistic picture of the influx for every activity, because the table only explain how many activities that are occurring in the different care pathways, but not how many activities that are actually happening. However, the table yields some indications on the distribution. The coloring of each entry is a function of two parameters, the number of times the activity occurs in the associated care pathway, and how often the activity is carried out overall in all the care pathways. Therefore, a more reddish color tells us that the activity is included many times in the associated care pathway, and it also included to a great extent in total across the different care pathways. This coloring is used in order to identify critical activities which is common for all the care pathways.

Table 19: Comparison of the number of activities in the different care pathways.

Activity	**Uterine**	**Cervical**	**Ovarian**
Referral	1	1	1
GI	3	1	0
Biopsy	1	0	1
CT	1	2	0
MRI	0	2	0
MDT	0	1	1
OC	2	1	4
Chemotrapy	1	2	3
Surgery	2	2	2
Radiotherpay	0	3	0
Brachyterapy	0	1	0
Blood sample	0	1	0

Figure 27, 28 and 29 show the average number of waiting days for each of the queues in the three cancer care pathways. The associated activities for the queues are found in Table 7 and 8. Each bar in the bar chart represents the average waiting time for the associated queue. The bars are identified with a combination of letters and digits. The first letter identifies the care pathway, the second letter tells if it is a diagnosis (D) or a treatment (T) queue. The diagnosis queues only have one digit, which denotes the position of the queue in the associated diagnosis path. For the treatment queues, there are two digits. The first digit denotes the treatment path, and the second number represents the order of the queue in the associated path. The different treatment paths and related activities associated to each of the cancer care pathways are presented in Table 8. For example, CT-2-2, denotes the second queue in the second treatment pathway belonging to the cervical cancer care pathway, where the associated activity is an MDT meeting. The queues belonging to the diagnosis stage and the different treatment paths are given different colors in order to make them distinguishable.

The average waiting times for the queues in the uterine care pathway (Figure 27) are mostly evenly distributed, except for the queue UD-4 with an average waiting time of 7.2 days. The patients in this queue waits for a CT scan. As Table 19 shows, CT resources are distributed between the uterine and the cervical care pathway. The color state that the activity is performed a moderate number of times in the different care pathways. Furthermore, the resources associated with a CT scan are radiologists, radiographers and a CT scanner. Of these resources, the CT scanners and radiologists are the most limited. Because these resources are limited and used in several pathways, there is a high pressure on these resources which explains why the UD-4 queue has a high average waiting days.

On the other hand, the UD-3 queue has an average waiting time of 0.7 days. The associated activity is a biopsy. As Table 19 shows, there are less biopsy activities compared to the CT scan, in the different care pathways. The other biopsy queue, OD-2 (Figure 29), also has a low waiting time with an average of 1 day. The low waiting times for these queues are a result of the low number of queues, and because the associated resources have much capacity. In a broad view, it is positive with low waiting times. On the other hand, it may also indicate that the amount of available resources are not utilized in the most efficient manner.

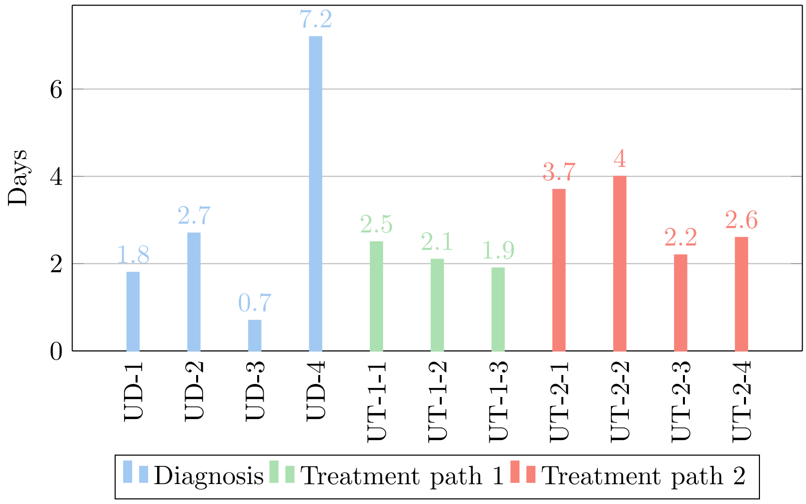


Figure 27: Average waiting days per queue in the uterine care pathway.

Earlier in this section, we have seen that the cervical cancer care pathway has both the highest number of patients residing in the associated queues, and it has also the highest average waiting time. The average waiting time per queue in the cervical cancer care pathway (Figure 28) shows that the diagnosis queues contribute the most to the long waiting times for patients following this care pathway. The diagnosis activities for the cervical cancer patients comprise of referral, a visit at the outpatient clinic, CT scan, MRI scan and a gynecological examination. These activities entail numerous resources that are used by all of the care pathways, which contribute to explain why cervical cancer patients are experiencing high waiting times during the diagnosis stage. With this in mind, increasing the performance of the diagnosis queues for the cervical care pathway may yield lower waiting times overall. The other queues do not vary much in waiting times, with an average between one and five days.

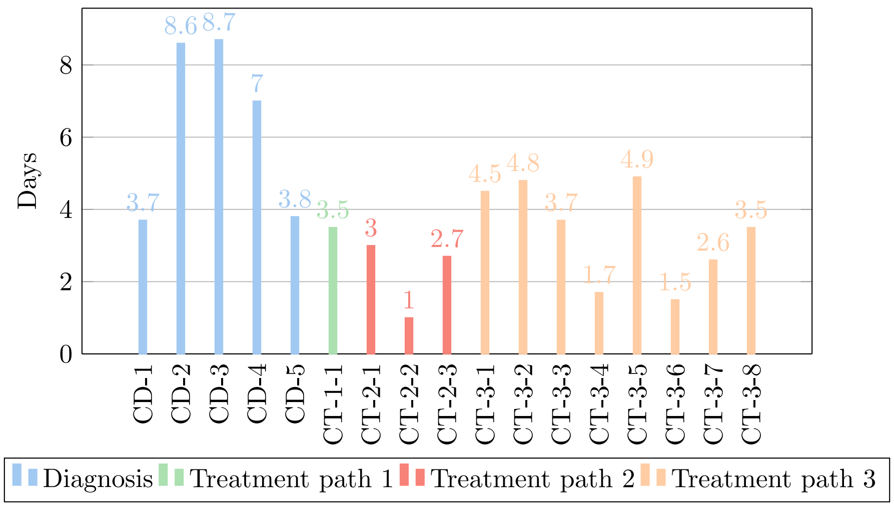


Figure 28: Average waiting days per queue in the cervical care pathway.

Figure 29, shows the average number of days waited for each queue in the ovarian care pathway. All of the queues, except for OT-2-3, have low average waiting times. The associated activity for the OT-2-3 queue is surgery. From Table 19, we observe that surgery is common in all of the pathways, contributing to more pressure on the resources. This explains one aspect of the high waiting time. Another reason for why the waiting times are building up in queue OT-2-3, is because most of the patients are following this pathway. From Table 13, we see that 65 % of the ovarian cancer patients end up in this path. We also observe that the third treatment path of the cervical care pathway receives 60 % of the cervical cancer patients, but this care pathway has less patients entering the pathway because the expected value for arrivals is lower.

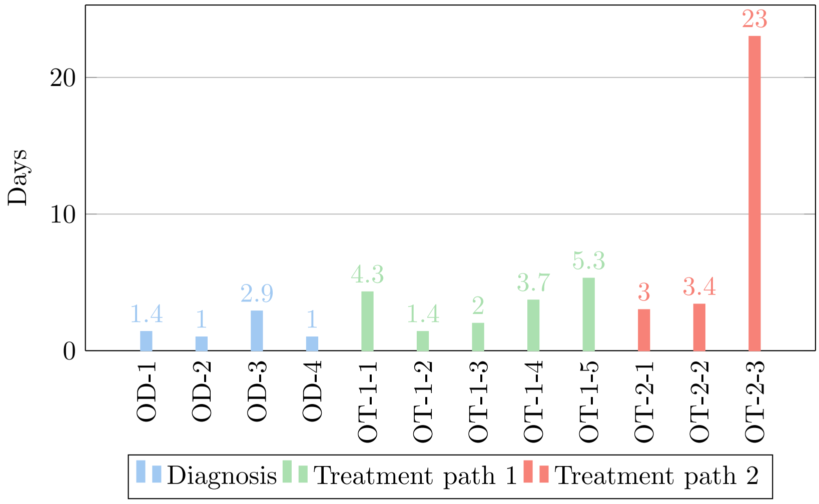


Figure 29: Average waiting time per queue in the ovarian care pathway.

Time limits

As described in Chapter 4, there are time limits associated to different phases of the diagnosis stage. Figure 30 shows the cumulative distribution of time limit violations throughout the simulation period, compared to the development of patients in the system. For example, the value for the phase 2 violations at day 200, means that 10 % of the associated violations happened up until this day. Thus a flat line implies that there are no violations occurring in the corresponding time interval. The lines seem to grow constantly, but that does not imply that they will grow over the value they have on the last day, if the simulation had continued. In a cumulative distribution of the share of patients who violates the deadline, when more patients leave the system, we get a larger value in the denominator, which shifts the graph further down if there are no more time limit violations. As observed in Figure 30, the distributions of time limit violations starts after day 49, which indicates the end of the warm-up period.

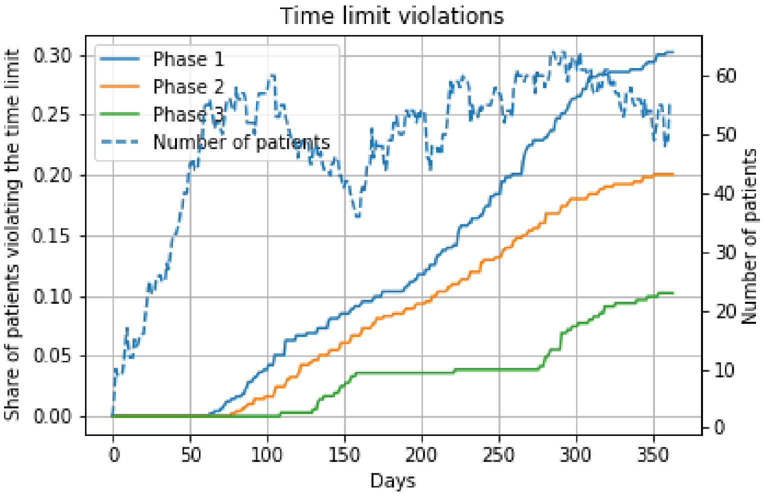


Figure 30: The share of time limit violations in the different phases. The dashed line represents the total number of patients in the system.

The phase 1 limit is of 6 days, and it denotes the time from referral until the first visit at the hospital. The phase 2 limit is 16 days, and it represents the time limit from first visit at the hospital until the last activity in the diagnosis stage is undergone. The phase 3 limit is 14 days, and it represents the time limit from the last activity in the diagnosis stage until the start of the first activity in the treatment stage. As it shows, the share of time limit violations for phase 1, 2 and 3 are 30 %, 20 % and 10 %, respectively.

In the case of the phase 1 time limit violation, it shows that the share of violations rise in a somewhat constant manner, only with small reductions in violations when the number of patients fall, as it does around day 150. This tells us that the phase 1 time limit is a limit that is difficult to comply with independent on how many patients that are residing in the system, and in total 30 % of the patients do not make it to the first visit at the hospital within the time limit.

Figure 30 shows that the share of phase 2 violations are evenly increasing, but to a lesser extent compared to the phase 1 violations. It also shows that the line is independent of the fluctuations of the number of patients in the system. As with the phase 1 limit, the phase 2 limit is also difficult to comply with independent of the number of patients residing in the system.

The green line that represents the phase 3 violations, tends to rise when the number of patients reaches an amount of about 55, as is the case at day 100 and day 280. Furthermore, the graph is virtually flat, between day 150 and 280, telling us that there are no time limit violations within this time interval. Therefore, the phase 3 violations are to a greater extent more dependent on the number of patients in the system, compared to the two other time limits. A reason for this may be a combination of two aspects. Firstly, there are no activities between the ending of the diagnosis stage and the beginning of the treatment stage, this reduce the possibilities of accumulating more waiting time. The other aspect, is that the phase 3 time limit more relaxed than the others, given the number of activities in between.

The time limits associated to each of the cancer types are the same in the case study. This alone would motivate a behavior where the waiting times for each of the three cancer types were more or less the same. However, the care process for cervical cancer patients has more activities than the two others, which in turn increases the waiting time for these patients. The number of activities associated to a care process should not alone determine how much waiting time a patient should expect - in other words, cancer patients with complex diseases should not be punished with longer waiting times just because their care processes require more devotion from the hospital.

It must be pointed out that the graph only shows if there has been a violation of the time limit or not, it does not tell anything about by how much the time limit is surpassed. Thus, the time limits may not be used as a sole measure of the performance in the different phases of the diagnosis stage.

Resource utilization

Figure 31 shows the resource utilization rate for each week that is simulated. From the figure, we observe that the utilization rate oscillates between 60 % and 80 %, but tends to converge towards a rate between 70 % and 80 %. Furthermore, if we compare the resource utilization with the number of patients in the system, illustrated in Figure 24, we see that the utilization rate follows the development of the number of patients in the system. This tells us that when many patients reside in the system, the utilization of the resources is higher.

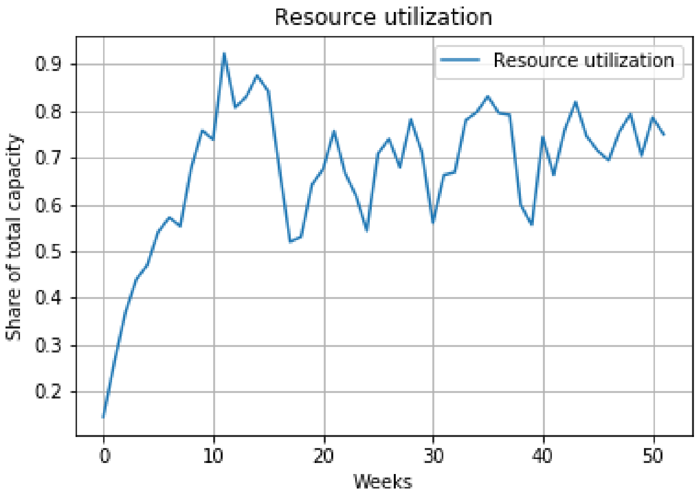


Figure 31: Resource utilization for each week in the simulation period.

9.3.2 Changing the scheduling policy

Until this point, the scheduling policy in the scheduling framework has been based on choosing the patient that has been waiting the most throughout her time in the system. In this section, we look into the effects of changing to a first in, first out (FIFO) policy, where patients that enter a queue first are serviced first.

In order to change the policy, the optimization model and the simulation model both need to be adjusted. For the optimization model, the objective function weights are set to penalize patients with high values of the n-index, therefore, the new objective weights are no longer dependent on the number of days waited in the system m as was the case for the original policy. Since the objective weights only are dependent on the n index, we get, $W_n = n$, giving the objective function shown in Equation (21). The new objective function motivates the optimization model to schedule services for the patients who have been in the same queue the longest. This is in contrast to the original objective function, where the patients who have the highest accumulated waiting time, m, is chosen. When it comes to the simulation model, when a service event happens, the model chooses the patient having the earliest arrival timestamp.

$$\min z = \sum_{j \in \mathcal{J}} \sum_{t \in \mathcal{T}} \sum_{n \in \mathcal{E}^N} \sum_{m \in \mathcal{E}^M} n q_{jtnm} \tag{21}$$

Figure 32 shows the development of the number of patients residing in the system when using the new scheduling policy. If we compare this development to the one in Figure 24 we see that the average number of patients in the system is larger when the FIFO policy is applied. In Figure 24 the number of patients oscillates between 40 and 60 after the warm-up, whereas in Figure 32 the number of patients has in average increased, and it stabilizes between 50 and 70. The reason for the increase of patients in the system comes from the fact that when the optimization model uses the FIFO policy, its focus is no longer to reduce the amount of total waiting periods for the patients, but rather to reduce the waiting time associated with the current queue the patient is waiting in. This further causes patients to reside in the system for a longer time, thus the number of patients is increased.

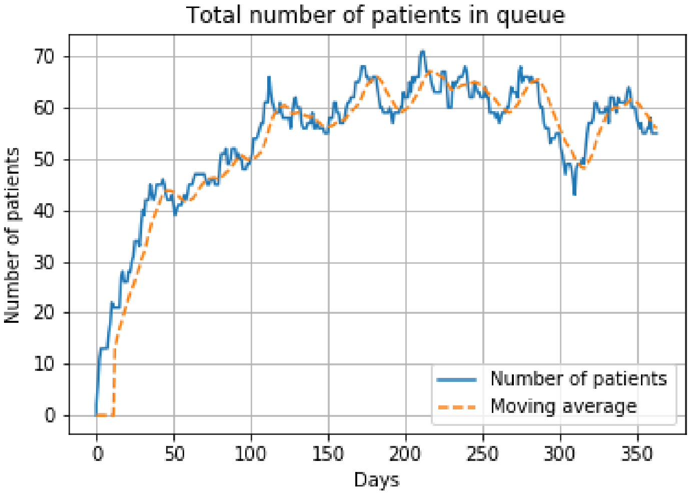


Figure 32: Number of patients in the system when the FIFO policy is applied.

In Figure 33, the development of the number of patients in the different care pathways is shown. Comparing this graph with the one in Figure 25 shows that all of the cancer care pathways have an increased number of patients in their associated queues, and the increase is distributed somewhat evenly over the different pathways.

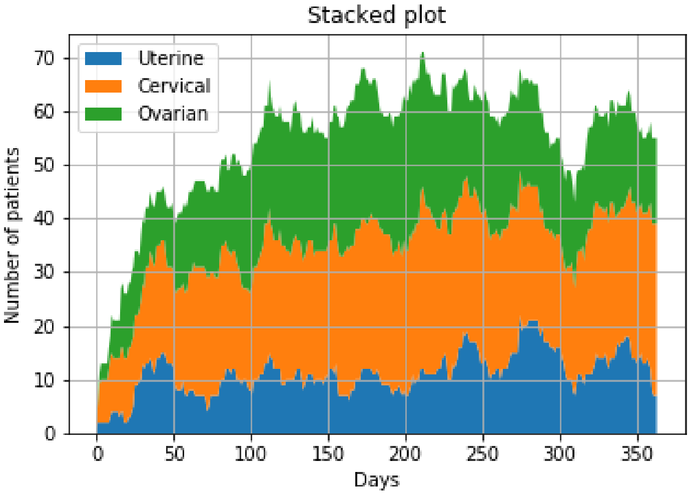


Figure 33: Number of patients in the system per patient group when the FIFO policy is applied.

The change in scheduling policy's impact on waiting times

Figure 34 compares the average waiting times for the original policy and the FIFO policy. The overall average waiting time has increased from 34 in the original policy to 37 for the FIFO policy. The main drivers for this increase are the uterine and ovarian care pathways. As seen in the figure, the average waiting time for the uterine care pathway has increased with 5 days, and the average waiting time for the ovarian care pathway has increased with 6 days. As discussed above, the FIFO policy results in an increased number of patients in the system. This further implies longer queues, and thus also longer average waiting times for most of the patients.

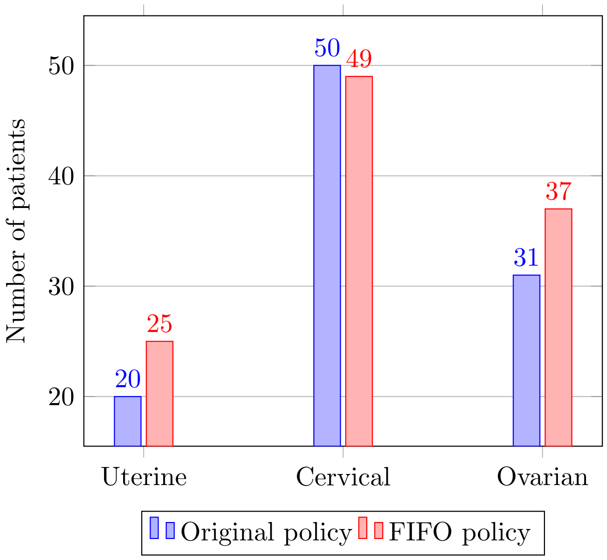
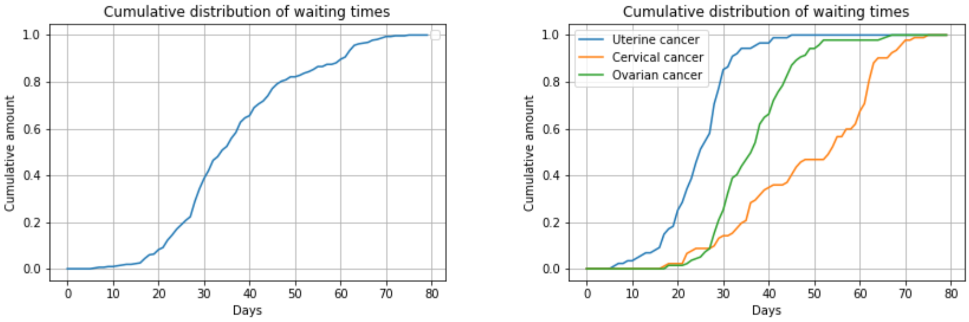


Figure 34: Comparison of the waiting times for the different care pathways for different scheduling policies.

Figure 35a shows the cumulative distribution of waiting times for the FIFO policy. The graph shows that the cumulative distribution associated with the FIFO policy is shifted more to the right compared to the distribution in Figure 26, implying longer waiting times. In the original case, 50 % of the patients wait for 30 days or less, while in the case of the FIFO policy, only 40 % of the patients wait for 30 days or less. Figure 35b shows the cumulative distribution of waiting times for the different cancer care pathways. It shows that the distributions for the uterine and ovarian pathway are shifted to the right, implying longer waiting times. The distribution for the cervical care is more flattened out, but still quite similar to the original policy as expected considering the fairly similar waiting times.



(a) Cumulative distribution - all patients aggregated.

(b) Cumulative distribution - per cancer type.

Figure 35: Cumulative distribution of waiting times when applying the FIFO policy.

9.3.3 Reflections on the results

In the introduction to this thesis, we addressed the issue of possible pitfalls related to organizational silos that hospitals may fall into. We advocate an approach where these organizational silos do not stand in the way for efficient planning and use of resources. As discussed in Section 9.3.2, it was found that changing the scheduling policy to FIFO in a queue, did increase the overall waiting times for patients compared to our main policy, which is to assign queue positions based on how long the patients have waited in total. Assigning positions in a queue using a FIFO policy may encourage a practice of silo mentality. Here, one looks for the best solution in the short term, not taking into account interdependence in the system or how a decision affects the total waiting time. Of course, we have not proven mathematically that the scheduling policy we use is better in terms of waiting time, in general, than the FIFO policy. However, it is noted that in our case, where we to a larger extent have a system-wide approach to planning, turned out to be a more efficient scheduling policy when the overall goal is to minimize patient waiting times. On the other hand, we do not deny the fact that there exist situations where the FIFO policy is the most suitable scheduling policy.

In every research project, this included, there are ethical considerations that must be made. For example, we need to consider any possible conflict of interest. In a hospital setting, such a conflict may arise in the process of resource allocation. As described in the introduction to this thesis, different hospital departments may have different reward schemes or different objectives. Our work attempts to neglect this type of conflicts, and strive for an integrated approach to planning, where all decisions are made in order to fulfill the overall goal of reducing patient waiting times and not to fulfill the objective of one hospital department in isolation from the rest of the hospital. Another ethical consideration that arises in the intersection between operations research and health care, is prioritization. Is it appropriate for a mathematical program to prioritize patients solely based on mathematical input, or should this process be based on human judgment? Generic scheduling policies are efficient and do as they are told. But, as patients get comorbidities and have more complex health conditions, human intelligence might be more important than ever before in planning decisions in health care.

10 Concluding remarks

This Master's thesis has introduced the problem of tactical patient admission and resource planning for patients enrolled in a standardized care process. We present a mathematical formulation describing the problem, and introduce a scheduling framework where the results from the optimization model are simulated and evaluated taking uncertainty into account. With this approach, we are able to assess the dynamics of the system over multiple planning periods. To solve the optimization problem, a MIP solver was utilized.

To evaluate the work, the optimization and simulation models work in an iterative process, referred to as the scheduling framework, using a rolling horizon approach. The optimization model solves the problem for one planning period in a deterministic setting. The solution for parts of the planning period is then exposed to a stochastic and realistic environment in the simulation model. The simulation is here only regarded as a substitute for an actual implementation of the solution from the optimization model. In the stochastic simulation environment, uncertainties are managed by drawing random events from probability distributions. We manage uncertainties associated with the number of patients referred to the hospital each day, which treatment a patient receives after a diagnosis is set, how much time each patient will use at each activity and if a patient shows up to her activity or not on time.

The optimization model may serve as a decision-support tool for hospital planners facing planning problems similar to the one presented in this thesis. The scheduling framework is able to evaluate solutions that adhere to relevant restrictions in a realistic setting. We conduct a computational study where we test how the system behaves when parameters in the optimization model and simulation model are altered.

In the technical study of the optimization model, it becomes evident that we are able to solve the problem inspired by the Oslo University Hospital with three different, standardized cancer care pathways. We are able to solve the case for planning periods up 28 days to optimality in reasonable time. However, the problem size escalates rapidly when the planning period increases, and for the UCO instance with a planning period of 35 days, the optimization model could not find the optimal solution within the time limit of 10,800 seconds.

The computational study reveals that waiting times vary when altering simulation parameters such as the planning period, implementation period, rate of no-shows and shifts from one iteration to the next in the rolling horizon approach. The results show that increasing the length of planning period, slightly shifts the distribution of waiting times so that the waiting times are decreased for the majority of the patients. It is also shown using a practical example that waiting times tend to decrease when the implementation period is shorter. A disadvantage of shorter implementation periods is less flexibility for the hospital staff. We also show that a scheduling policy where patients are serviced based on their total waiting time is preferred over a FIFO scheduling policy for each individual queue.

In the case study performed on data inspired by Oslo University Hospital, we are able to admit patients and allocate resources in a manner that shows satisfying results in terms of waiting times. The results show that there are variations in the waiting times among the patients residing to the uterine, cervical and ovarian cancer care pathways. The difference revealed, where cervical cancer patients experience the longest waiting times, is expected due to their more extensive diagnosis stage compared to the other two cancer types. We also evaluate how resources are utilized and asses the model's ability to adhere to cancer type-specific time limits.

Using the scheduling framework, alternative schedules can readily be evaluated by the hospital planners. Planning today is a to a large distinct done manually. By employing the scheduling framework, we are able to quantify key parameters needed to solve the problem. However, planning in hospitals in general is an extensive and complex process. Consequently, accurate input data from the hospital is essential to obtain successful results from the model, and supplementary discussions with Oslo University Hospital regarding the input data and following adjustments could make the results more accurate.

The findings in this Master's thesis may assist hospital management in the planning process for patients enrolled in a standardized care process. In general, the decision maker should have a broad understanding of the practical problem at hand when using the work of this thesis in considerations of how its solutions acts in a real life situation. Optimal solutions from the optimization model evaluated and endorsed by the simulation model are not automatically the best solution in a real life setting, and they can not take into account all aspects of the problem, such as human considerations and assessments. We believe that by mixing human insights and expertise with our analytical approach, an overall satisfying solution can be obtained.

11 Future research

In this thesis, the problem of tactical patient admission and resource planning for patients enrolled in a standardized care process in a multi-disciplinary care system is solved. Here, resource capacities are allocated among hospital activities for patients enrolled in a standardized care process in a multi-disciplinary care. The optimization model have proven successful in fulfilling in finding an alternative way of generating schedules. In the following sections, we suggest areas of research that may be continued on as an extension to our work.

11.1 Uncertainty in resource capacities

Resource capacity is subject to a high degree of uncertainty. One thing is generating a schedule for resources, another thing is to be able to execute it. The problem of uncertain access to resources is especially evident in multi-disciplinary care systems where the sequence of appointments is of high importance, i.e. if the first appointment in a flow-shop type of system is canceled for a patient, the entire series of appointments for the patient must be re-scheduled. We encourage future researchers to handle resource capacities stochastically. This challenge is relevant for problems similar to the one presented in thesis, since information regarding resource capacities are revealed in the time between the moment of decision-making and the actual moment to which the decision applies.

11.2 Cooperation and user-friendliness

Although the mathematical model and the proposed scheduling framework are the most vital elements of this thesis, implementability is also of uttermost importance. An optimal solution to a mathematical problem is not necessarily equivalent to the optimal solution to the practical problem that is being modeled. It is of highest important that future research on OR methods in multi-disciplinary planning is carried out in close cooperation the hospital and its key personnel that are to actualize the model. One should also in the future highlight the user-friendliness of the implementation, by e.g. developing a graphical user interface that eases the implementation and does not require the personnel implementing the model to have extensive knowledge of OR.

11.3 The full care process

In this thesis, the patient cancer care process is modeled only for the diagnosis and treatment stages. These are key elements in a cancer care process, but do not fully depict the real situation of the problem. An important aspect of health care, and cancer care in particular, is the follow-up and control of patients that have undergone treatment. In future research, it is of interest to assess additional stages of the care process to enable an even closer integration of activities in the care process. There are potential great benefits, and challenges, of increased integration. A trade-off between on one hand providing an extensive system scope and on the other hand providing enough details of the system, must be done. Incorporating both views to the fullest extent is not possible due to issues with complexity. By extending the scope of the problem, more variables and constraints are added to the problem, and solving the optimization model would become harder. Therefore, if this extension is to be added, an alternative solution method, for example heuristics (Section 11.5), is preferred.

11.4 Similar areas of application

Although our work contributes to the planning of a multi-disciplinary health care system where the system is assumed to be a hospital, our work is applicable to other aspects of health care. The insights that we provide for a cancer clinic is transferable to other types of health care, where initial tests and examinations of the patient are decisive in the treatment of the patient, e.g. rehabilitation after a major surgery.

11.5 Alternative solution procedure

The optimization model is solved using a MIP solver. When encountering instances with multiple cancer care processes over multiple weeks, the runtime for the model increases significantly, and an optimal solution is not to be found. To find optimal, or near optimal solutions to the optimization problem, we suggest using heuristics. It may be difficult to prove optimality when using heuristic, but a good and feasible plan is also valuable in a hospital planning setting. A good heuristic is able to exploit the physical interpretation of the problem better than an exact solution procedure, as the one presented in this thesis. As introduced in the literature review in Chapter 3, a genetic algorithm is an alternative solution procedure that has been successfully implemented in settings similar to the setting of this thesis.

11.6 Multi-objective optimization

Evaluating multiple objective functions simultaneously would increase the practical use of the model. In a hospital planning problem, there will always be different stakeholders with different demands and requirements. Patients wish to minimize their waiting time, the hospital wants to utilize its resources efficiently as possible, doctors ask for a work-life balance, tax-payers want their money to be used sensibly and emergency patients need access to care without having to wait. It is of course impossible to please all interested groups simultaneously. In our objective function, we mainly focus on the patients, by minimizing patient waiting time. However, an interesting aspect would be to extend the objective function with an objective explicitly representing an objective of the doctors. This would better reflect the situation at the hospital, where decisions must be balanced to meet the needs and requirements of different stakeholders. This would in turn make the problem harder to solve, since a procedure to rank solutions is needed.

12 Appendices

Appendix A Literature review summary table

Paper	Application area	Objective function	Solution method	Type of system	Variability approach
Our contribution	Cancer care clinic	Min waiting times	MIP solver, evaluated using simulations	Mixed shop	Deterministic
Alfonso et al. (2011)	Blood collection system	-	Discrete event simulation	Flow shop	Stochastic patient behavior
Azadeh et al. (2014)	Emergency patient care	Min waiting time for patients	Genetic algorithm and simulation	Open shop	Deterministic
Azadeh et al. (2015)	Care processes without a patient present	Min total completion time	Genetic algorithm and simulation	Flow shop	Deterministic
Barz et al. (2015)	Elective patient care	Max expected contribution	Markov decision process, solved using ADP and heuristics	Flow shop	Stochastic care pathways
Bikker et al. (2015)	Radiotherapy	Balance supply and demand	IP model evaluated using discrete-event simulation	Flow shop	Deterministic
Cardoen et al. (2008)	Elective patient care	-	Discrete event simulation	Flow shop	Stochastic patient arrivals, appointment durations and care pathways
Castro et al. (2012)	Radiotherapy	Min number of patients exceeding waiting time targets	Hierarchical, mixed integer programming	Flow shop	Deterministic
Chern et al. (2008)	Outpatient and day care clinics	Min waiting time of doctors and patients	Health Examination Scheduling Algorithm (HESA)	Mixed shop	Deterministic
Conforti et al. (2010)	Elective patient care	Max number of admitted patients	Exact solution of MIP model	Mixed shop	Deterministic
Dharmadhikari et al. (2011)	Outpatient and day care clinics	-	Heuristic block scheduling with different priority schemes	Open shop	Stochastic patient arrivals
Du (2013)	Clinical pathways scheduling	Min makespan	Hybrid genetic algorithm	Flow shop	Deterministic
Gartner et al. (2014)	Planning of patient flow	Min makespan	Exact solution of MIP model	Flow shop	Stochastic patient arrivals
Hahn-Goldberg et al. (2014)	Outpatient and day care clinics	Min makespan	Constraint programming	Mixed shop	Stochastic patient arrivals
Hulshof et al. (2013)	Clinical pathways scheduling	Min number of waiting patients	Exact solution of MIP model	Flow shop	Deterministic
Hulshof et al. (2015)	Clinical pathways scheduling	Min patient waiting times	Approximate dynamic programming	Flow shop	Stochastic patient arrivals and care pathways
Jeric et al. (2011)	Outpatient and day care clinics	Min waiting times and makespan	Variable neighbourhood search	Flow shop	Deterministic
Jeric et al. (2012)	Clinical pathways scheduling	Multi-objective	Metaheuristics (VNS, SS, NSGA)	Open shop	Deterministic
Leeftink et al. (2019)	Outpatient and day care clinics	Multi-objective	Sample average approximation algorithm	Flow shop	Stochastic care pathways
Liang et al. (2015)	Chemotherapy	Min patient waiting times and max throughput	Discrete event simulation	Flow shop	Stochastic patient arrivals and appointment durations
Matta et al. (2007)	Cancer care clinic	Multi-objective	Discrete event simulation	Mixed shop	Deterministic
Perez et al. (2013)	Nuclear medicine	-	Scheduling algorithms	Flow shop	Stochastic patient arrivals and appointment durations
Petrovic et al. (2012)	Radiotherapy	Min patient waiting time	Genetic algorithm	Flow shop	Stochastic patient arrivals
Romero et al. (2012)	Cancer care clinic	Min throughput time	Discrete event simulation	Flow shop	Deterministic
Saadani et al. (2014)	Outpatient and day care clinics	Min stay durations	Exact solution of MILP model	Mixed shop	Deterministic
Sadki et al. (2011)	Chemotherapy	Min waiting times and makespan	Lagrangian relaxation	Flow shop	Stochastic arrivals
Saremi et al. (2015)	Outpatient and day care clinic	Min waiting times and makespan	Multi-agent tabu search algorithm	Mixed shop	Stochastic service times
Vermeulen et al. (2007)	Outpatient and day care clinics	Min completion time of all patients	Pareto-improvement appointment exchanging algorithm	Open shop	Deterministic
Vrugt et al. (2017)	Cancer care clinic	-	Simulation	Flow shop	Stochastic patient arrivals and appointment durations

Appendix B Alternative procedure for determining the objective function weights

In the following, an alternative iterative approach for determining the objective function weights for the optimization model is presented. This iterative approach is inspired by the algorithm presented in Hulshof et al. (2013). The overall goal of the procedure is for the objective weights to be set in such a way that they align with the different performance targets the hospital may have on the different queues. The procedure was not used because it did not satisfy our goal of reducing the overall waiting times in a sufficient manner. Furthermore, to assign the target values for the different queues also proves to be challenging. Firstly, some new terminology is introduced, thereafter the procedure is presented. Finally, results when running the procedure are evaluated. New parameters used in the procedure are introduced in Table 20.

Table 20: Additional parameters used in the procedure.

Parameter	**Description**
C_{jt}	Access time for a queue j at time t
C_{jt}^T	Target value for the access time for a queue j at time t
C_{jt}^P	Performance measure for the access time of a queue j at time t
α	The percentile of a queue that is accounted for in order when the access time is calculated.
B_{jt}^T	Target value for number of patients serviced in queue j at time t.
B_{jt}^P	Service performance measure for a queue j at time t
θ	The value that decides the strictness of the convergence criterion for the procedure
ϵ	Small number used when the objective weights are calculated

One of the performance measures that is used in the procedure is based on the access time, C_{jt}, of a queue, defined in Equation (22). The access time of a queue at a given time is the number of time periods the α percentile of the patients have been waiting. For example, if α equals 0.5 for a given queue j at time period t, the access time is the minimal number of time periods that ensure 50% of the patients are serviced within this time period.

$$C_{jt} = \min \left\{ n | \sum_{n=0}^{n} \sum_{m \in \mathcal{E}} q_{jtnm} > \alpha \sum_{n=0}^{\mathcal{E}^N} \sum_{m \in \mathcal{E}} q_{jtnm} \right\} \qquad \forall j \in \mathcal{J}, t \in \mathcal{T} \quad (22)$$

Furthermore, the hospital may set target values, C_{jt}^T, for the access times associated with the different queues at different time periods. By dividing the access time by the target value for the access time, we get the level of performance for the given queue at the given time period, as presented in Equation (23). A value that exceeds one, implies that the queue is underperforming considering the target value. The queue is underperforming because it has a higher access time compared to the target of the hospital. On the other hand, if the value is below one, the queue is overperforming.

$$C_{jt}^P = \frac{C_{jt}}{C_{jt}^T} \quad (23)$$

The hospital may also assign target values to the different queues denoting the number of patients to be serviced. Furthermore, if the target value for the number of patients serviced at queue j in time period t is divided by the number of patients that is actually serviced in the same queue and time period, we get the level of service performance, given in Equation (24). If the service performance has a value of more than 1, the queue is underperforming because there are less services than the target value. While, if the score is below 1, the queue is overperforming.

$$B^P_{jt} = \frac{B^T_{jt}}{b_{jt}} \tag{24}$$

The objective weights are defined in Equation (25). We see that the weights are a function of a parameter s, which denote the current iteration. The equation shows that if a patient has waited for zero time periods, the associated weight gets a value of zero. If the value of n is more than zero, the weight is determined by the variables $u_j(s)$ and $m_j(s)$. These variables are defined in Equation (26) and (27), where the $m_j(s)$ variables are raised to n. In this manner, the weights motivate the model, in an increasing way, to service patients residing in queues with a high n index.

$$W_{jn}(s) = \begin{cases} 0, & \text{if } n = 0 \\ u_j(s)m_j^n(s), & \text{if } n > 0 \end{cases} \qquad \forall j \in \mathcal{J} \tag{25}$$

The procedure

The procedure follows three steps in order to update the W_{jn} variables.

Step 1:

In step 1, $W_{jn}(s)$ is initialized with the following values: $s = 0$, $u_j(0) = 1$ and $m_j(0) = 1 + \epsilon$, where ϵ is a small number. The ϵ is needed in order to ensure that u_j is greater than zero, and m_j to be greater than one. If the ϵ is not implemented, it is not given that the values for the objective function weights increase with the n index. Finally, using the initial objective weights, the MIP is solved.

Step 2:

In step 2, the iteration counter is first incremented, $s = s + 1$, and new values for $u_j(s)$ and $m_j(s)$ are calculated based on the result from the MIP, using Equation (26) and (27).

$$u_j(s) = \max\left\{0 + \epsilon, u_j(s-1) + \frac{1}{s}\sum_{t=0}^{T} B^P_{jt} - 1\right\} \qquad \forall j \in \mathcal{J} \tag{26}$$

$$m_j(s) = \max\left\{1 + \epsilon, m_j(s-1) + \frac{1}{s}\sum_{t=0}^{T} C^P_{jt} - 1\right\} \qquad \forall j \in \mathcal{J} \tag{27}$$

If the result from the sum after the substraction in Equations (26) and (27) is negative, the queue is overperforming, i.e. more resource capacities than required are allocated to this queue. This oveperformance is handled by decreasing the parameters u and m. On the contrary, if we get a positive number after the subtraction, the queue is underperforming and the parameters are thus increased. The performance measure is a measure for all of the time periods, because it is summarized over all time periods. The new values for $u_j(s)$ and $m_j(s)$ are then used to update $W_{jn}(s)$. Thereafter, the MIP is solved again with the new weights.

Step 3:

If the convergence criteria shown in Equation (28) is met, the iteration stops. This convergence criteria is invoked if the difference in value for $u_j(s)$ and $m_j(s)$ between two consecutive iterations

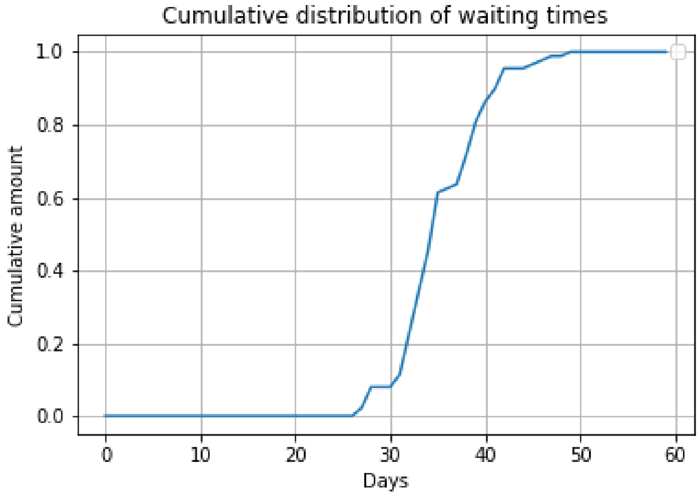


Figure 36: Cumulative distribution of waiting times using an iterative approach for determining the objective function weights

are less than a parameter θ. If the criterion is not met, the procedure continues and steps 2 and 3 are repeated.

$$\max\{|u_j(s) - u_j(s-1)|, |m_j(s) - m_j(s-1)\} \leq \theta \qquad \forall j \in \mathcal{J} \quad (28)$$

Results

The procedure must be run before every call to the optimization model in the scheduling framework (Chapter 7). This is because the weights adjust to the different performance measures which may change from one iteration to the next. Incorporating the procedure in the scheduling framework makes every iteration a comprehensive and time consuming process, because the framework has to run the optimization model several more times in order to make sure that the objective weights converge in every iteration. The procedure was tested on the instance U-14, described in Table 15. Furthermore, the service and access time targets were assigned values which favored the queues belonging to the diagnosis phase, because these are the queues that have the most flow of patients.

The cumulative distribution of waiting times, resulting from the run of the scheduling framework, is shown in Figure 36. We observe that the distribution of waiting times are shifted strongly to the right, compared to the distributions presented in Section 9.2, with an average waiting time of 35 days. It shows that the procedure does not give any improvement of the waiting times in this setting. Additionally, for the procedure to give any reasonable results the target values should be determined by hospital staff, which may be a difficult job, and the results should also be evaluated by the same staff.

Appendix C Resource overview

Table 21: Activities and the associated resources.

Activity	Resources
Referral	None
Gynecological examination	Gynecologist, nurse, outpatient clinic room
Biopsy	Gynecologist, pathologist, laboratory
CT	Radiologist, radiographer, CT scanner
MRI	Radiologist, radiographer, MRI scanner
MDT	Gynecologist, physician, meeting room
Outpatient clinic	Physician, nurse, outpatient clinic room
Chemoterapy	Physician, nurse, day unit
Surgery	Physician, nurse, operating room
Radiotherapy	Gynecologist, radiographer, radiotherapy room
Brachyterapy	Radiologist, radiographer, radiotherapy room
Blood sample	Physician, outpatient clinic room

Table 22: Resource capacities.

Activity	Capacity (hours per day)
Physician	48
Gynecologist	56
Radiologist	4
Radiographer	4
Pathologist	4
Nurse	48
CT scanner	2
MRI scanner	2
Operating room	16
Laboratory (Biopsy)	4
Outpatient clinic	4
Day unit	16
Radiotherapy room	4
Meeting room	16

References

Aarskog, H. and Lindstad, J. (2019). Optimization-based block planning of CT scanners and radiologists at Radiumhospitalet. Project report. Norwegian University of Science and Technology.

Addis, B., Carello, G., Grosso, A., and Tànfani, E. (2015). Operating room scheduling and rescheduling: a rolling horizon approach. *Flexible Services and Manufacturing Journal*, pages 206–232.

Afroze, T. and Gardell, M. R. (2015). Algorithm construction for efficient scheduling of advanced health care at home. `https://pdfs.semanticscholar.org/c035/4cadecacaa874f66897aec5c6134b7b5cf55.pdf`. Accessed April 29, 2020.

Ahmad, O. B., Boschi-Pinto, C., Lopez, A. D., Murray, C. J. L., Lozano, R., and Inoue, M. (2001). Age standardization of rates: A new WHO standard. Technical report, World Health Organization. GPE Discussion Paper Series: No.31.

Alfonso, E., Xie, X., Augusto, V., and Garraud, O. (2011). Modeling and simulation of blood collection systems. *Health Care Management Science*, 15:63–78.

Allemani, C., Matsuda, T., Di Carlo, V., Harewood, R., Matz, M., Niksic, M., Bonaventure, A., Valkov, M., Johnson, C. J., Estève, J., Ogunbiyi, O. J., e Silva, G. A., Chen, W.-Q., Eser, S., Engholm, G., Stiller, C. A., Monnereau, A., Woods, R. R., Visser, O., Lim, G., Aitken, J., Weir, H. K., and Coleman, M. P. (2018). Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37,513,025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *The Lancet*, 391(10125):1023–1075.

Azadeh, A., Baghersad, M., Fahrani, M. H., and Zarrin, M. (2015). Semi-online patient scheduling in pathology laboratories. *Artificial Intelligence in Medicine*, 64:217–226.

Azadeh, A., Fahrani, M. H., Torabzadeh, S., and Bahersad, M. (2014). Scheduling prioritized patients in emergency department laboratories. *Computer Methods and Programs in Biomedicine*, 117:61–70.

Badar, M. A., Samidi, S., and Gardner, L. (2013). Reducing the bullwhip effect in the supply chain: A study of different ordering strategies. *The Journal of Technology Studies*, 39:52–63.

Bailey, N. T. J. (1952). A study of queues and appointment systems in hospital outpatient departments with special reference to waiting times. *Journal of the Statistical Society*, 14(2):185–199.

Barz, C. and Rajaram, K. (2015). Elective patient admission and scheduling under multiple resource constraints. *Production and Operations Management*, 24(12):1907–1930.

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

Cancer Registry of Norway (2019). Cancer in Norway 2018 - cancer incidence, mortality, survival and prevalence in Norway. `https://www.kreftregisteret.no/globalassets/cancer-in-norway/2018/cin2018.pdf`. Accessed April 20, 2020.

Cancer Registry of Norway (2020). The cancer registry's online statistics bank. `https://sb.kreftregisteret.no/insidens/`. Accessed February 26, 2020.

Cardoen, B. and Demeulemeester, E. (2009). Capacity of Clinical Pathways - A Strategy Multilevel Evaluation Tool. *Journal of Medical Systems*, 32:443–452.

Castro, E. and Petrovic, S. (2012). Combined mathematical programming and heuristics for a radiotherapy pre-treatment scheduling problem. *J Sched*, 15:333–346.

Chen, Z., King, W., Pearcey, R., Kerba, M., and Mackillop, W. J. (2008). The relationship between waiting time for radiotherapy and clinical outcomes: A systematic review of the literature. *Radiotherapy and Oncology*, 87(1):3–16.

Chern, C.-C., Chien, P.-S., and Chen, S.-Y. (2008). A heuristic algorithm for the hospital health examination scheduling problem. *European Journal of Operational Research*, 186:1137–1157.

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

Dharmadhikari, N. and Zhang, J. (2013). Simulation optimization of blocking appointment scheduling policies for multi-clinic appointments in centralized scheduling systems. *International Journal of Engineering and Innovative Technology*, 2(11).

Du, G., Jiang, Z., Yao, Y., and Diao, X. (2013). Clinical pathways scheduling using hybrid genetic algorithm. *J Med Syst*, 9945(37).

Figueira, G. and Almada-Lobo, B. (2014). Hybrid simulation-optimization methods: A taxonomy and discussion. *Simulation Modelling Practice and Theory*, 46:118–134.

Forrester, J. W. (1961). *Industrial dynamics*. M.I.T. Press, Cambridge, Massachusetts.

Gartner, D. and Kolisch, R. (2014). Scheduling the hospital-wide flow of elective patients. *European Journal of Operational Research*, 233(3):689–699.

Green, L. (2006). Queuing analysis in Healthcare. In Hall, R. W., editor, *Patient Flow: Reducing Delay in Healthcare Delivery*, volume 91 of *International Series in Operations Research & Management Science*, chapter 10, pages 281–307. Springer, Boston, MA.

Gullhav, A. N., Christiansen, M., Nygreen, B., Aarlott, M. M., Medhus, J. E., Skomsvoll, J., and Østbyhaug, P. O. (2018). Block scheduling at magnetic resonance imaging labs. *Operations Research for Health Care*, 18:52–64.

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

Hans, E. W., van Houdenhoven, M., and Hulshof, P. J. H. (2012). A framework for healthcare planning and control. In Hall, R., editor, *Handbook of Healthcare System Scheduling. International Series in Operations Research & Management*, volume 168, chapter 12, pages 303–320. Springer, Boston, MA.

Hansen, P., Mladenovic, N., Brimberg, J., and Peréz, J. A. M. (2019). Variable Neighborhood Search. In Gendreau, M. and Potvin, J. Y., editors, *Handbook of Metaheuristics*, International Series in Operations Research and Management Science, chapter 3, pages 61–86. Springer.

Hansen, R. P., Vedsted, P., Sokolowski, I., Søndergaard, J., and Olesen, F. (2011). Time intervals from first symptom to treatment of cancer: a cohort study of 2,212 newly diagnosed cancer patients. *BMC Health Services Research*, 11(284):1–8.

Helse- og omsorgsdepartementet (2013). Sammen mot kreft. nasjonal kreftstrategi 2013-2017. `https://www.regjeringen.no/no/dokumenter/sammen---mot-kreft/id728818/`. Accessed April 20, 2020.

Hillier, F. S. and Lieberman, G. J. (2015). *Introduction to Operations Research*. McGraw-Hill Education, New York, 10 edition.

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

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

Hulshof, P. J. H., Mes, M. R. K., Boucherie, R. J., and Hans, E. W. (2015). Patient admission planning using approximate dynamic programming. *Flexible Services and Manufacturing Journal*, 28(1).

Hurst, J. (2000). Challgenges for health systems in member countries of the organisation for economic co-operation and development. *Bulletin of the World Health Organization*, 78(6):751–760.

Jerić, S. V. and Figuiera, J. R. (2012). Multi-objective scheduling and a resource allocation problem in hospitals. *Journal of Scheduling*, 15:513–535.

Jerić, S. V., Pacheco, J. A., and Lukač, Z. (2011). Use of VNS heuristics for scheduling of patients in hospital. *Journal of the Operational Research Society*, 62:1227–1238.

Johannesen, T. B. (2014). Stor økning i krefttilfeller fram mot 2030. `https://www.kreftregisteret.no/Generelt/Nyheter/Stor-okning-i-krefttilfeller-fram-mot-2030/`. Accessed April 20, 2020.

Kalton, A., Singh, M., August, D., Parin, C., and Othman, E. (1997). Using simulation to improve the operational efficiency of a multi-disciplinary clinic. *Journal of the Society for Health Systems*, 5(3):43–62.

Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks*, volume 4, pages 1942–1948.

Kristensen, G. B., Lindemann, K., Solheim, O., and Bruheim, K. (2017). Oncolex - gynekologisk kreft. `http://oncolex.no/GYN`. Accessed April 24, 2020.

Leeftink, A. G., Boucherie, R., Hans, E. W., Verdaasdonk, M., Vliegen, I., and van Diest, P. (2016). Batch scheduling in the histopathology laboratory. *Flexible Services and Manufacturing Journal*, 30(31):171–197.

Leeftink, A. G., Vliegen, I. M. H., and Hans, E. W. (2019). Stochastic integer programming for multi-disciplinary outpatient clinic planning. *Health Care Management Science*, 22:53–67.

Leeftink, G., Bikker, I., Vliegen, I. M. H., and Boucherie, R. J. (2018). Multi-disciplinary planning in health care: a review. *Health Systems*, pages 1–24.

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

Mackillop, W. J. (2007). Killing time: The consequences of delays in radiotherapy. *Radiotherapy and Oncology*, 84:1–4.

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

Matta, M. E. and Patterson, S. S. (2007). Evaluating multiple performance measures across several dimensions at a multi-facility outpatient center. *Health Care Management Science*, 10:173–194.

Norwegian Directorate of Health (2020). Pakkeforløp for kreft. `https://helsenorge.no/sykdom/kreft/pakkeforlop-for-kreft`. Accessed April 22, 2020.

Norwegian Patient Registry (2020). Statistics from the Norwegian patient registry on cancer incidences. `https://www.helsedirektoratet.no/statistikk/statistikk-fra-npr/pakkeforlop-for-kreft-indikatorer-for-aktivitet-og-forlopstid`. Accessed April 22, 2020.

Oslo University Hospital (2019). Pakkeforløp og koordinatorfunksjonen. `https://oslo-universitetssykehus.no/fag-og-forskning/kvalitet/pakkeforlop-og-koordinatorfunksjonen`. Accessed May 6, 2020.

Petrovic, D., Castro, E., Petrovic, S., and Kapamara, T. (2013). Radiotherapy scheduling. In Etaner-Uyar, A. S., Özcan, E., and Urquhart, N., editors, *Automated Scheduling and Planning. Studies in Computation Intelligence*, chapter 7, pages 155–190. Springer-Verlag Berlin Heidelberg, Berlin.

Pham, D. N. and Klinkert, A. (2008). Surgical case scheduling as a generalized job shop scheduling problem. *European Journal of Operational Research*, 185(3):1011–1025.

Pérez, E., Ntaiomo, L., Wilhelm, W. E., Bailey, C. R., and McCormack, P. (2013). Patient and resource scheduling of multi-step medical procedures in nuclear medicine. *IIE Transactions on Healthcare Systems Engineering*, 1:1114–1136.

Richards, M. A. (2009). The national awareness and early diagnosis initiative in England: assembling the evidence. *British Journal of Cancer*, 101(2):1–4.

Risberg, T., Sørbye, S. W., Norum, J., and Wist, E. (1996). Diagnostic delay causes more psychological distress in female than in male cancer patients. *Anticancer Research*, 16(2):995–1000.

Romero, H. L., Dellaert, N. P., van der Geer, S., Frunt, M., Jansen-Vullers, M. H., and Krekels, G. A. M. (2012). Admission and capacity planning for the implementation of one-stop-shop in skin cancer treatment using simulation-based optimization. *Health Care Management Science*, 16:75–86.

Rutqvist, L. E. (2006). Waiting times for cancer patients - a ”slippery slope” in oncology. *Acta Oncologica*, 45(2):121–123.

Saadani, N. E. H., Bahroun, Z., and Bouras, A. (2014). A linear mathematical model for patients' activities scheduling on hospital resources. *2014 International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 74–80.

Sadki, A., Xie, X., and Chauvin, F. (2011). Appointment scheduling of oncology outpatients. *IEEE International Conference on Automation Science and Engineering.*

Saremi, A., Jula, P., El Mekkawy, T., and Wang, G. G. (2015). Bi-criteria appointment scheduling of patients with heterogeneous service sequences. *Expert Systems with Applications*, 42:4029–4041.

Sorensen, J. R., Johansen, J., Gano, L., Sørensen, J. A., Larsen, S. R., Andersen, P. B., Thomassen, A., and Godballe, C. (2014). A ”package solution” fast track program can reduce the diagnostic waiting time in head and neck cancer. *European Archives of Oto-Rhino-Laryngology*, 271:1163–1170.

Statistics Norway (2018). Norge – god helse og store utgifter. `https://www.ssb.no/helse/artikler-og-publikasjoner/norge-god-helse-og-store-utgifter`. Accessed November 29, 2019.

Tretlie, S. (2016). Dramatisk kreftøkning neste 15 år. `https://www.kreftregisteret.no/Generelt/Nyheter/kreft-i-2030/`. Accessed April 20, 2020.

United Nations (2019a). World population 1950-2020. `https://population.un.org/wpp/Download/Standard/Population/`. Accessed April, 2020.

United Nations (2019b). World population prediction. `https://population.un.org/wpp/Download/Probabilistic/Population/`. Accessed April 15, 2020.

Vanberkel, P. (2011). *Interacting hospital departments and uncertain patient flows: theoretical models and applications.* PhD thesis, University of Twente, Netherlands.

Vanberkel, P., Boucherie, R. J., Hans, E. W., and Johann L. Hurink, N. L. (2009). A survey of health care models that encompass multiple departments. *University of Twente, Department of Applied Mathematics*, 1.

Venkitasubramanian, A., Roberts, S. D., and Joines, J. A. (2015). Object oriented framework for healthcare simulation. In *Proceedings of the 2015 Winter Simulation Conference*, WSC '15, pages 1436–1446, Huntington Beach, California. IEEE Press.

Vermeulen, I., Bohte, S., Somefun, K., and Poutré, H. L. (2007). Multi-agent pareto appointment exchanging in hospital patient scheduling. *Service Oriented Computing and Applications*, 1:185–196.

von de Vrugt, M., Boucherie, R. J., Smilde, T. J., de Jong, M., and Bessems, M. (2017). Rapid diagnoses at the breast center of Jeroen Bosch hospital: a case study invoking queuing theory and discrete event simulation. *Health Systems*, 6:77–89.

Wang, X. and Disney, S. M. (2016). The bullwhip effect: progress, trends and directions. *European Journal of Operational Research*, 250(3):691–701.

WHO (2007). Cancer control: Prevention. WHO guide for effective programmes. `https://www.who.int/cancer/publications/cancer_control_prevention/en/`. Accessed April 20, 2020.

Zotteri, G. (2013). An empirical investigation on causes and effects of the bullwhip-effect: Evidence from the personal care sector. *International Journal of Production Economics*, 143(2):489–498.

NTNU
Norwegian University of
Science and Technology