Why wait?

Organizing integrated processes in cancer care

Gréanne Leeftink

WHY WAIT?

Organizing Integrated Processes in Cancer Care

Gréanne Leeftink

Dissertation committee

Chairman & secretary:	Prof. dr. T.A.J. Toonen *University of Twente, Enschede, the Netherlands*
Promotors:	Prof. dr. ir. E.W. Hans *University of Twente, Enschede, the Netherlands* Prof. dr. R.J. Boucherie *University of Twente, Enschede, the Netherlands* Dr. ir. I.M.H. Vliegen *Eindhoven University of Technology, Eindhoven, the Netherlands*
Members:	Prof. dr. N. Litvak *University of Twente, Enschede, the Netherlands* Dr. K.S. Pasupathy *Mayo Clinic, Rochester, MN, USA* Prof. dr. S. Siesling *University of Twente, Enschede, the Netherlands* Prof. dr. G.D. Valk *University Medical Center Utrecht, Utrecht, the Netherlands* Dr. W.B. de Vries *University Medical Center Utrecht, Utrecht, the Netherlands*

Ph.D. thesis, University of Twente, Enschede, the Netherlands
Beta Research School for Operations Management and Logistics
Center for Healthcare Operations Improvement and Research
Center for Telematics and Information Technology (No. 17-444, ISSN 1381-3617)

This work is part of the NWO Talent research program 'Rapid diagnostics for cancer? Yes we can!' with project number 406-14-128, which is financed by the Netherlands Organisation for Scientific Research (NWO).

Printed by: Ipskamp Printing, Enschede, the Netherlands
Cover design: Christel Haitsma-Wolters, Enschede, the Netherlands
Cover portrait photo: Gijs van Ouwerkerk Fotografie, Enschede, the Netherlands

ISBN 978-90-365-4411-5
DOI 10.3990/1.9789036544115

WHY WAIT?

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
Prof. dr. T.T.M. Palstra,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 15 december 2017 om 16:45 uur

door

Anne Greetje Leeftink

geboren op 30 augustus 1991
te Smallingerland, Nederland

Dit proefschrift is goedgekeurd door de promotoren:

Prof. dr. ir. E.W. Hans
Prof. dr. R.J. Boucherie

Iubilate Deo

Voorwoord

Als eindopdracht van het bachelor honoursprogramma in 2011 moest ik een onderzoeksvoorstel schrijven. Ik kon me niets saaiers voorstellen op dat moment, en probeerde er maar het beste van te maken. Terugkijkend is het de meest invloedrijke opdracht geweest die ik heb gehad in mijn studie. Ik had nooit kunnen beseffen dat ik 6,5 jaar later, met een beurs op zak en veel ziekenhuiservaring rijker, op datzelfde onderzoeksvoorstel zou promoveren.

Deze thesis was er niet geweest zonder de hulp en inzet van heel veel mensen. Zonder tekort te doen aan mijn dankbaarheid voor anderen, wil ik een aantal mensen in het bijzonder bedanken.

Allereerst wil ik mijn promotor, Erwin, bedanken. Jij hebt me weten uit te dagen om steeds een stapje verder te gaan, van het schrijven van het betreffende (Nederlandstalige) onderzoeksvoorstel, via Canada, naar dit promotieonderzoek op een NWO-voorstel. Het werk dat je nu in handen hebt was er niet gekomen zonder jouw gedrevenheid en enthousiasme. Ook de manier waarop je omgaat met je studenten is inspirerend en heeft me veel geleerd over het begeleiden van studenten. Dankjewel dat jouw deur altijd voor mij openstaat, niet alleen om nieuwe onderzoeksideeen uit te wisselen, maar ook voor een oneindige aanvoer van koffie en leuke liedjes voor de gitaar.

Richard, jouw kritische blik, scherpe feedback en directe aanpak hebben me geholpen mijn werk te verbeteren. Je hebt me geleerd om het overzicht te behouden, focus aan te brengen en de grote lijnen van mijn onderzoek goed uit te denken. Bedankt daarvoor.

Ingrid, helaas hebben wij mijn promotietraject niet samen kunnen afronden, maar zonder jou was het begin heel anders verlopen. Bedankt voor de vele waardevolle gesprekken, zowel werkgerelateerd als op persoonlijk vlak. Jij wist mij te focussen, als ik me in al mijn enthousiasme veel te veel op de hals haalde. Ik hoop dat je heel gelukkig bent en blijft met jouw gezin en baan in Eindhoven.

I would like to thank my committee members, Nelly Litvak, Kalyan Pasupathy, Sabine Siesling, Gerlof Valk, and Willem de Vries, for your time and valuable feedback regarding my thesis and PhD defense.

Daarnaast ben ik dank verschuldigd aan de vele mensen met wie ik samen heb gewerkt aan de hoofdstukken van dit proefschrift. Onze samenwerking is essentieel geweest in de uitvoering van het onderzoek en zonder jullie was dat niet gelukt.

Voor Hoofdstuk 1 bedank ik Linda, Zeno en Gijs. Ondanks dat jullie afstuderen niet tot een publicatie heeft geleid, heb ik veel geleerd van ons project.

Ook bedank ik Sabine, Janine, Ingrid, Els, Maarten, Jelle en Richard voor hun bijdragen aan Hoofdstuk 1.

Ingeborg, dankjewel voor de succesvolle samenwerking bij ons literatuuronderzoek (Hoofdstuk 2). Het was leuk om in het buitenland zo toch nog een beetje binding te houden met de UT, en, terwijl ik weer terug was, op de hoogte te blijven van jouw belevenissen in Canada.

Marina, Paul en alle collega's bij de Pathologie, hartelijk bedankt voor de samenwerking en het vertrouwen tijdens en na mijn afstudeerproject. Dankzij jullie voortvarendheid, enthousiasme en kennis zijn Hoofdstuk 3 en 4 mogelijk gemaakt.

I thank Kal, Mustafa, Esra, and Gabriela for facilitating my stay in the USA and for their contributions to Chapter 5.

Hanneke en de HPB-collega's, bedankt voor jullie bijdrage aan Hoofdstuk 6.

Martijn, dankjewel voor je schijnbaar oneindige enthousiasme voor Hoofdstuk 7. Ik zie ernaar uit om hier nog samen verder aan door te werken. Astrid, Bart en Floor, bedankt voor de hulp in de strijd om data. Het is op de valreep toch maar mooi gelukt!

Erwin, ik ben er trots op dat ons oneindige project nu afgerond en gepubliceerd is, en een mooi 8^{e} hoofdstuk van mijn proefschrift mag vormen. De lange doorlooptijd heeft het onderzoek in mijn ogen sterker gemaakt.

Ik heb gedurende mijn promotietraject het voorrecht gehad een groot aantal studenten te mogen begeleiden: Frank, Jurre, Simone, Linda, Elieke, Rianne, Myrthe en Karlijn, Zeno en Gijs, Thijmen, Wouter, Bryan, Marleen, Panagiotis, Matthew, Joran, Wouter, Robin, Pleuni, Kelvin, Eelco, Koen, Marjanne, Anneloes, Robert, Davey, Arjan, Benjamin en Cynthia. Jullie inzet en onderzoek is de grootste impact in de zorgpraktijk (Hoofdstuk 9). Dank voor jullie bijdragen. Ik heb veel van jullie geleerd en ik hoop dat jullie er met net zoveel plezier op terugkijken als ik.

In de afgelopen drie jaar heeft een enorm aantal collega's in Utrecht, Enschede en Rochester een bijdrage geleverd aan mijn promotietraject, zowel in de vorm van een directe bijdrage aan mijn proefschrift, als indirect, door goede gesprekken, wijze raad, en gezelligheid tijdens lunches, koffiepauzes en congressen.

Ik heb met veel plezier bij het UMC Utrecht, en in het bijzonder bij het Cancer Center mogen werken. Alle collega's wil ik dan ook bedanken voor het vertrouwen, de mogelijkheden en gezelligheid van de afgelopen jaren. In het speciaal wil ik een aantal mensen noemen.

Jos en Bert, bedankt dat jullie mijn promotieplek in het UMC Utrecht gefaciliteerd hebben en mij de mogelijkheden hebben geboden om overal mee te kijken en mee te denken. Jos, de spreuk die ik van je kreeg bij jouw afscheid was een schot in de roos; hij heeft het tot een van mijn stellingen geschopt. Hanneke, dankjewel dat je mijn rechterhersenhelft wilde aanvullen. Ik hoop dat ik met mijn linkerhersenhelft ook jou heb mogen verrijken. Floor, samen hebben we een aantal mooie overwinningen behaald en kunnen bijdragen aan betere patiëntenzorg. Ik vind het gaaf om te zien hoe jij altijd doorzet, dankjewel voor de samenwerking. Astrid en Dick, bedankt voor de vele datasets die jullie mij

keer op keer hebben aangeleverd. Dit heeft mijn onderzoek een stuk gemakkelijker gemaakt. Arjan, bedankt voor de vele brainstormsessies en het vinden van nieuwe uitdagingen in de opstartfase van mijn onderzoek. Miranda, ik heb veel geleerd van jouw vastberadenheid en je ambitie. Dankjewel dat je me na jouw vertrek uit het Cancer Center niet uit het oog verloor. Michel, jouw enthousiasme en inzichten om capaciteitsmanagement op de kaart te zetten in het UMC Utrecht werken aanstekelijk. Dankjewel voor de tijd die je voor mij nam om mijn ideeën te spiegelen, en zelfs delen van mijn proefschrift door te lezen. Ik hoop dat we nog een lange en vruchtvolle samenwerking tegemoet gaan. Last but not least mijn kamergenoten op Q4. Bedankt voor jullie gezelligheid en de goede gesprekken. Jullie hebben ervoor gezorgd dat ik me altijd welkom voelde.

Ook op de Universiteit Twente zijn er een heleboel mensen die mijn promotietraject een stuk leuker en makkelijker hebben gemaakt. Mijn collega's en kamergenoten bij IEBIS en SOR (en DWMP) wil ik bedanken voor de afgelopen jaren. Ik vond het een voorrecht om deel te mogen uitmaken van twee vakgroepen.

I thank my office mates at IEBIS, among who Abhishta, Andrej, Floor, Lucas, Nardo, Sjoerd, and Wouter, for all the times they made me coffee, and for helping me out with all kind of questions. Elke, dank voor je enthousiasme, de goede gesprekken met wijze raad en het vele lachen. Jij weet alles voor elkaar te krijgen en jouw aanwezigheid maakt het werk een stuk leuker.

Daarnaast wil ik mijn collega's en kamergenoten van CHOIR in het bijzonder bedanken. Het is gaaf om deel te zijn van een groep onderzoekers die in een vergelijkbare omgeving onderzoek doet en mede daardoor elkaar en elkaars onderzoek kunnen versterken. Naast van de inhoud, heb ik ook erg genoten van de gezelligheid op de vrijdagen en tijdens lunchwandelingen, uitjes en barbecues, en de spelletjes en uitstapjes tijdens conferenties. Aleida, Maartje en Nardo, bedankt voor de goede voorbeelden die jullie mij hebben gegegven. Jullie promotietrajecten en proefschriften zijn een inspiratiebron voor mij geweest en ik zie ernaar uit om jullie nog regelmatig op de UT tegen te blijven komen. Ingeborg, ik zal onze lunch in Nashville niet snel vergeten. Joost en Sem, het was leuk om samen met jullie in Spanje rond te reizen voorafgaand aan ORAHS 2016; Ik weet nog steeds hoe ik koffie voor ons drietjes moet bestellen. Sem, het is gaaf dat we samen gaan promoveren, ik hoop dat het een mooi feest wordt! Ingeborg, Joost, Thomas, Bruno, Shiya, Maarten, Jasper en Eline, heel veel succes met de afronding van jullie PhD.

I feel honored that I got the opportunity to spend part of my PhD in the USA in Rochester's Mayo Clinic. Kal and Mustafa, thank you for facilitating me as a visiting researcher in the fall of 2016. I am proud that our joint work resulted in a chapter in my thesis, I thank you for your collaboration and time to make this work, and look forward to future collaborations. Gabriela, a special thanks for all the time we spent together. I felt very welcome, and really enjoyed our coffee breaks together.

Mijn bezoek aan de Verenigde Staten was niet mogelijk geweest zonder de financiële ondersteuning van het Data-Piet Fonds van het Prins Bernhard Cul-

tuurfonds, hartelijk bedankt daarvoor.

Tot slot wil ik mijn man, familie en vrienden bedanken. Jullie steun, gezelligheid en afleiding was meer dan welkom. Christel, ik vind het gaaf dat je naast onze trouwkaart, nu ook mijn proefschrift hebt ontworpen. Je bent een topper! Marinke, de afgelopen jaren was jij degene die mijn promotie en alle keuzes die daarbij komen kijken het allerbeste begreep. Je wist zelfs je onverwachte bezoekje aan ons klushuis perfect te timen! En onze ministudie, met n=2, laat zien dat een huis kopen en promoveren fantastisch goed samengaat! Dankjewel voor je vriendschap, je luisterend oor, de gezellige etentjes en dat je mijn paranimf wil zijn. Marjella, ook in jouw leven zijn mooie dingen gebeurd de afgelopen jaren, en ik vind het leuk dat we daardoor weer wat dichter bij elkaar in de buurt wonen. Dankjewel dat je mijn paranimf wil zijn. En maak je daar geen zorgen over, het is net zoals op een AV, maar dan wel bij een studentikoze vereniging.

Papa, mama, Hanneke, Caroline en Stein, en Geert, ik vind het gaaf dat ik jullie er als familie bij heb gekregen. Bedankt dat er altijd een bord eten en een schoon bed voor me klaarstond wanneer ik veel in Utrecht moest zijn.

Pap en mam, er is teveel om jullie voor te bedanken. Jullie hebben mij grootgebracht tot wie ik nu ben. Jullie hebben mij de kansen gegeven mijn talenten te ontwikkelen en mij daarin altijd uitgedaagd en gesteund. En tegelijkertijd helpen jullie me altijd te beseffen wat echt belangrijk is in het leven. Karine en Sjors, Marjella en Wicher, en Gerbert, bedankt voor de gezellige weekenden en luisterende oren. Jullie zijn stuk voor stuk fantastische personen, en zonder jullie zou het leven een stuk saaier zijn geweest.

Lieve Dirk, wat hebben wij de afgelopen jaren veel meegemaakt samen. We zijn getrouwd, we hebben een huis gekocht en zijn elkaar na een paar maanden klussen nog steeds niet zat! Ik geniet erg van ons leven samen, en ik hoop dat God ons nog vele jaren samen zal geven. Jouw realisme en liefdevolle steun zorgen ervoor dat ik met beide benen op de grond blijf staan, ook als ik mijzelf (en jou...) in mijn enthousiasme te veel op de hals haal. Dankjewel dat je me altijd hebt gesteund in mijn promotietraject, ook toen ik besloot om je voor een aantal maanden in Nederland achter te laten. Ik houd van je!

God, ik dank U voor al deze lieve mensen om mij heen, voor de gaven en talenten waarmee U mij zegent, en voor de mogelijkheden die U mij de afgelopen jaren heeft gegeven. Alle eer aan U!

Gréanne
Apeldoorn, november 2017

Contents

PART 1

introduction

Why Wait?
Organizing Integrated Processes in Cancer Care

Chapter 1

E. Visser, A.G. Leeftink, P.S.N. van Rossum, S. Siesling, R. van Hillegersberg, and J.P. Ruurda. Waiting time from diagnosis to treatment has no impact on survival in patients with esophageal cancer. *Annals of Surgical Oncology*, 23(8):2679-2689, 2016.

E. Visser, P.S.N. van Rossum, A.G. Leeftink, S. Siesling, R. van Hillegersberg, and J.P. Ruurda. Impact of diagnosis-to-treatment waiting time on survival in esophageal cancer patients A population-based study in The Netherlands. *European Journal of Surgical Oncology*, 43(2):461-470, 2017.

Chapter 2

A.G. Leeftink, I.A. Bikker, I.M.H. Vliegen, and R.J. Boucherie. Multi-disciplinary appointment planning - a review. *Submitted.*

CHAPTER 1

Motivation of this work

1.1 Introduction

Being diagnosed with cancer can be devastating. Cancer has been the number one cause of death in the Netherlands since 2008, and one third of the Dutch population will be diagnosed with cancer during their life [57].

In the Netherlands, patients receive a high quality of care [204]. Research shows that successful treatment plans exist for many different types of cancer [278]. However, patients might not receive this care to their full advantage if they have to wait for their care.

Together with a growing demand of care, cost of care is growing, especially in cancer care [204]. The capacity is limited and resources are scarce. The aging population is causing the labor population to decrease over the coming decades. It is a challenge to improve health care processes with the existing resources [287].

Currently, the access time to diagnosis is several days to weeks, highly varying upon cancer type. Thereafter, a patient might need to wait several weeks before the treatment can actually start, due to waiting lists for the various treatment modalities. Patients can clinically deteriorate due to tumor growth during the waiting time. However, not all tumors have the same growth rate. A rapid diagnosis and treatment is highly recommended from a medical perspective for some specific cancer types such as breast and lung cancer [8], while for others waiting time is less likely to influence the patient's survival probability [201, 278].

Aside from the tumor growth rate, waiting influences the patient's psychosocial well-being. The emotional impact grows every day, when someone suspects to have cancer [231]. Therefore we need to strive to limit the time until a diagnosis is confirmed (along with a treatment plan) and start the care pathway as soon as possible.

Fast-track care pathways, such as rapid diagnostics, are not new within oncology. It has already been used to reduce the length of multiple care pathways, such as for breast cancer diagnostics. However, performance is typically not equitably divided among all patients. In the design of fast-track care pathways, there is a tradeoff between the economies of focus and economies of scale. On the one hand there are increased economies of focus for a specific patient population, which include providing more efficient care through standardization, and a possible improved quality of care through specialization. On the other hand, the

economies of scale reduce for the remaining remaining (complex) care, through reduced flexibility. Research shows that reserving capacity for patients with a specific care pathway results in a large waiting time increase for the other patients [297, 326].

The research in this thesis aims to improve the quality and efficiency of cancer care processes, by realizing quick access to care for all patients using existing resources. We develop new planning and control approaches to optimize the organization of multiple shared resources involved, so that access to diagnostics and treatment is equally divided and optimized over all patient types. We analyze and validate these through mathematical modeling and simulation.

To ensure practical relevance of this research, we intensively collaborate with the Utrecht Cancer Center department of University Medical Center Utrecht (UMC Utrecht), a large academic hospital in the Netherlands. This allows for close involvement of clinicians in the research and improvement projects, enables us to gather real-life data and problems, and provides a first user for implementing prototype outcomes.

The remainder of this chapter is organized as follows. First, Section 1.2 introduces Operations Management in healthcare. Section 1.3 describes the processes involved in the diagnostics and treatment of a cancer patient, followed by Sections 1.4 and 1.5 that show the relevance of delivering timely care from a medical and patient point of view respectively. Section 1.6 discusses cancer care from a hospital perspective, and Section 1.7 continues with a description of UMC Utrecht. We end with Section 1.8, which gives an outline of the remainder of this thesis.

1.2 Healthcare Operations Management

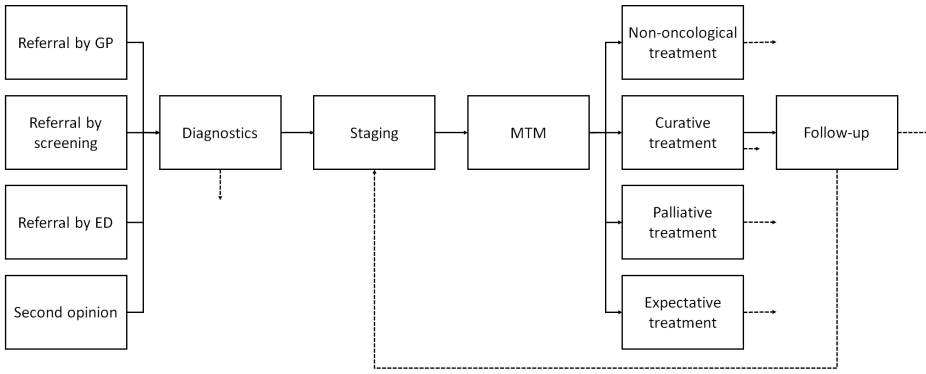
Process optimization in a healthcare context using a scientific approach is studied in the field of Operations Management (OM) in healthcare. OM entails the design, planning and control, and optimization of the organization of processes, and focuses on making them as efficient and effective as possible [279]. To aid decision-making in the design of these processes, we make use of Operations Research (OR) techniques, which include algorithmic optimization, modeling and evaluation methodologies such as mathematical programming, queueing theory, and computer simulation.

The impact of multiple changes to the healthcare system can be prospectively assessed without (possibly negatively) interfering in practice using OM and OR models. This way, these techniques can support decision makers in healthcare by developing (near-)optimal solutions, and by analyzing and evaluating the consequences of possible interventions to the healthcare system. We distinguish the following model categories:

Deterministic models include (integer) linear programming models. A linear programming model finds an optimal solution given an objective and set of constraints, which are all formulated using linear functions and equations.

Heuristics procedures can be subdivided in constructive heuristics, and im-

Figure 1.1 Care pathway of a cancer patient



provement heuristics. Constructive heuristics, such as greedy procedure, design a solution from scratch, whereas improvement heuristics, such as simulated annealing procedures, genetic algorithms, and local search approaches, try to improve upon a given solution. Although heuristic procedures not necessarily result in the optimal solution, they are designed to find an approximate optimal solution within reasonable time.

Simulation models are virtual models that represent a real world system. These models are used to analyze the performance of real world systems, and to experiment with possible interventions.

Stochastic models include markov models, queueing models, and stochastic analytical approaches, such as stochastic programming and robust optimization models. These models have in common that they incorporate some level of uncertainty, for example by incorporating random variables in the model formulation.

OM and OR methodologies are used in many applications, including production, process, and service industries. Examples of applications that we will refer to in this thesis are among others the process flow in the chemical industry [122, 130, 203], airline management [24], vehicle routing [160, 280], and timetabling [247]. Although already in the 1950's OR techniques were used to improve healthcare processes [20, 319], only recently OM and OR researchers gave considerable attention to the optimization of healthcare processes. Applications in healthcare include outpatient appointment scheduling, surgery scheduling, and nurse rostering. For an extensive overview of OM/OR healthcare applications we refer to Hulshof et al. [145].

1.3 Cancer diagnostics and treatment

The care pathway of a cancer patient, also referred to as the patient journey or clinical course, is displayed in Figure 1.1, and will be further explained in the text below.

Although not covered in this thesis, the patient journey through cancer care

most often starts outside the hospital, for example in a screening program, or when visiting a general practitioner (GP) [204]. Alternatively, patients might be redirected from the emergency department (ED) or from another hospital for a second opinion. As other researchers within the University of Twente are looking into these processes, these processes are out of the scope of this thesis.

When a patient suspects to have cancer, he or she enters a diagnostics trajectory. This can be a rapid diagnostic pathway, for example when suspecting to have breast cancer, but also a regular diagnostics trajectory. Rapid diagnostics has the advantage that multiple tests and consultations are scheduled on the same day, with dedicated staff. This ensures that a diagnosis can often be presented the same day. Within a regular trajectory, this might take more time, as appointment series are not aligned. The advantage of a regular trajectory on the other hand is that a patient gets more time to adjust to the idea of possibly having cancer. The final diagnosis is always confirmed by a pathologist, who examines a tissue sample on including possible tumorous cells. Other tests that are common are blood tests, endoscopic tests, and imaging. More specific details about the diagnostic trajectory of cancer patients can be found in Chapter 3, 4 and 6.

When initial diagnostics detects a tumor, follow-up diagnostic tests may be required to analyze the staging of the tumor in order to develop a treatment plan. For specific tumor categories, a patient can be referred to a specialized oncology center, for a one day visit to confirm the diagnosis and to offer a multidisciplinary approach in order to provide the treatment plan. The choice of a treatment modality depends on the type, size and location of the tumor, and of the patient characteristics and the stage of the disease. The results are discussed in a multi-disciplinary team meeting (MTM), in which a broad range of specialists and nursing staff gathers to discuss the treatment opportunities for the patient. Typically, an MTM takes place once per week, or once per day, depending on the tumor population.

Following the MTM, the patient and the designated treating clinician agree on the treatment plan. More specific details about this step can be found in Chapter 7. There are several treatment categories:

Curative treatment aims to cure the patient. This often involves surgery, to remove the tumor. Chemotherapy or radiation therapy can precede the surgery to reduce the size of the tumor (neoadjuvant therapy). Furthermore, such therapies can also be given after the surgery (adjuvant therapy), or in isolation, to minimize the risk of the cancer to recur. Other treatment modalities are for example hormonal therapy, and immunotherapy.

Palliative treatment aims to improve the quality of life for patients with no curative intent [149]. This might include relieving symptoms and side effects of curative treatment, although in this thesis, when referring to palliative treatment, we specifically refer to the care for patients with incurable cancer. Treatment includes pain relief and symptom control, together with attention for the psychological and spiritual needs of a patient.

Expectant treatment aims to maintain the state of the tumor, without taking

any direct action to treat the cancer, unless changes to the tumor occur or symptoms appear. For low-risk tumors, such as mantle cell lymphoma or prostate cancer, this provides an improved survival and a high quality of life for patients, who otherwise would have suffered from the side-effects of the other treatment modalities [73, 229]. Patients undergoing expectant treatment often receive a considerable amount of tests and exams, to closely watch the behavior of the tumor. This type of treatment is also known as deferred treatment.

Non-oncological treatment is required for patients who do not suffer from cancer. After the diagnostic trajectory, a patient may turn out to be cancer-free, or to suffer from another disease. In the latter case, patients might require non-oncological treatment.

In this thesis, we focus on curative treatment processes, with a focus on treatment including surgical removal of the tumor. Other treatment modalities, such as radiation therapy, are covered by other University of Twente researchers.

Following curative treatment, patients enter the follow-up, in which they are monitored for several years. After this period, patients are considered cured, and will leave the care system. However, during a check-up consultation in the follow-up, a recurring or new tumor might be found. In that case, after the required diagnostic tests, the patient will be discussed in the MTM again, followed by appropriate treatment.

Using an OM perspective, we need to consider the overall patient journey when optimizing cancer care processes. Patients may benefit from rapid diagnostics, but when they subsequently have to wait multiple weeks to start treatment, the advantages of rapid diagnostics quickly diminish. Therefore, Stichting Oncologische Samenwerking (SONCOS), a national umbrella organization for professionals and patients in cancer care, has set waiting time targets for every stage in the patient journey [282]. After referral by the GP, an appointment with a specialist should take place within one week. Furthermore, within 3 weeks after the first consultation in the hospital, a treatment plan has to be proposed to the patient, and within 6 weeks after the first consultation in the hospital this treatment should be started. Not only SONCOS, but also Dutch Cancer Society (KWF kankerbestrijding), a Dutch organization that focuses on cancer research, cancer policy, and cancer knowledge sharing, set waiting time targets for cancer care [154]. A patient should get access to a GP within 3 working days, and referral from the GP to the hospital cannot exceed 5 working days. Within 10 working days after the first consultation in the hospital a treatment plan has to be proposed, and within 15 working days after the treatment plan is known the treatment should start. Furthermore, they note that the total treatment time, including chemotherapy and radiation therapy if applicable, should be as short as possible.

Despite the various waiting time targets, hospitals still struggle with their waiting list management. In 2012, Netherlands institute for health service research (NIVEL) showed that 43% of the patients got access to the hospital within 5 working days, and only 54% of the patients were diagnosed within 10 working days.

In the following sections we will elaborate on the impact of waiting time in cancer care from a medical perspective (Section 1.4), from a patient perspective (Section 1.5), and from the perspective of the hospital (Section 1.6). This shows the urgency and main driver of reducing the waiting times in cancer care, and of cohering with the normative waiting times set by SONCOS and KWF.

1.4 Medical relevance

In the care process of a cancer patient, waiting time is considered as an important quality indicator [71]. Waiting time can occur before the first hospital visit (also known as access time), and between subsequent visits. During each waiting period, the tumor can grow, which might negatively affect the probability of disease free survival. However, the literature on the relationship between in-hospital waiting time and outcomes in cancer care is inconclusive. A relation between the waiting time from diagnostics to treatment and patient survival is reported for for example breast cancer [200, 258], head and neck cancer [133], and uterine cancer [95]. However, this relation is not proven for patients with several other cancer types, such as esophageal cancer [310, 311], lung cancer [219], and colorectal cancer [214, 252]. Causes for these mixed results can be differences in patient delay, due to the aggressiveness of the tumor, early or late manifestation of symptoms, and the presence or absence of screening programs [311].

Note that the aforementioned studies only include in-hospital waiting time (the time between diagnosis and start of treatment). The total waiting time also includes patient waiting time (time between onset of the symptoms and presentation to the GP), and the doctor waiting time (time between presentation to the GP and diagnosis) [310]. These periods may account for a larger amount of time of the total waiting time, as the patient delay can add up to several months to years, and thus influencing the patient's survival.

1.5 Patient relevance

Despite the lack of medical evidence for the waiting time as a prognostic factor for survival, multiple studies showed that waiting periods before treatment are distressing for patients and do seriously impact their quality of life. Patients prefer to be treated as soon as possible for the fear of tumor progression [70, 115, 139, 151, 209, 253, 312].

Organizing rapid diagnosis and treatment may come at a cost. In a pilot study, three University of Twente students analyzed what concessions patients are willing to make to get their diagnosis in one day and start treatment as soon as possible [110, 300]. Their studies show that patients are willing to travel longer. They are also willing to be served by multiple specialists, instead of one specialist that is dedicated to their case. Furthermore, they are willing to have as many examinations as needed in one day, if this all leads to lower waiting times. Furthermore, the patient's age is an important factor for the willingness to make

concessions for a rapid diagnosis and treatment. For example, younger patients were more likely to travel further than elderly patients. The recent trend of centralizing oncological care is in line with patient preferences, as centralization comes at the cost of longer travel times for patients, but offers hospitals the opportunity to organize their care more efficiently through economies of scale, which potentially results in lower access and waiting times.

1.6 Hospital relevance

Not only from a patient perspective, but also from a hospital perspective it is important to organize cancer care in an effective and efficient way.

An effective healthcare organization ensures the expected outcomes are reached; to cure as many (cancer) patients as possible, or, if this is not possible, to maximize their remaining (prolonged) quality of life. An efficient healthcare organization implies that hospitals use their resources in such a way that as many patients as possible receive the care they need. Efficiency increases the sustainability of the organization, as it allows more patients to be served. Especially in cancer care, it is important that healthcare institutes treat a considerable number of patients, in order to meet the national volume criteria. These criteria ensure that an institute has enough experience with treating a certain type of cancer in order to deliver a high quality of care.

Besides volume criteria, hospitals also face access time criteria, as mentioned in Section 1.3. Nowadays, patients are more informed and more selective in choosing a healthcare provider, and evaluate them regarding both quality outcomes and their waiting time performance.

Oncological care in the Netherlands is facing many changes in the near future from an organizational perspective. The number of patients with cancer is increasing, and treatment is becoming more complex and tailored to the needs of patients, which requires highly specialized specialists. This requires hospitals to join forces, and form comprehensive cancer networks on a regional and even national level [225]. Within such a network, care for patients with low-volume, complex diseases is centralized, in order to guarantee clinical expertise. The treatment of patients with high-volume diseases requires a well organized pathway as well, involving the specialists from multiple organizations. Although treatment is centralized, patients might want to have their follow-up and after-care closer to home. This requires a mentality shift for hospitals, towards patient-centered care. Chapter 6 and 7 present examples of how cancer care involving multiple specialists from several hospitals can be organized in a patient-centered way.

1.7 University Medical Center Utrecht

This thesis is the result of a collaboration between UMC Utrecht Cancer Center, and the Center for Healthcare Operations Improvement and Research (CHOIR) of the University of Twente. UMC Utrecht Cancer Center is to a large extent

the main driver behind the research presented in this thesis. Their ambition to continuously improve cancer care, and the questions raised by their staff, inspired most of the projects in this thesis.

UMC Utrecht is a large teaching hospital, which considers care, research, and education as their main tasks. UMC Utrecht approximately serves 30,000 inpatients and 100,000 outpatients, has 1,000 registered beds, and 11,000 employees (8,500 fte).

UMC Utrecht has identified six focus areas for the near future. One of them is 'cancer'. To this end, the Utrecht Cancer Center, one of UMC Utrecht's departments, aims to provide the best possible care for cancer patients, through, among others, multi-disciplinary collaborations and an excellent infrastructure [294]. This aim not only requires high quality care, but also high quality care delivery. Therefore, in their 'zorgconcept', a pamphlet in which they make promises to their patients regarding the care they aim to deliver, they state that patients can expect excellently organized care [58]. This requires efficient planning, tailored to the needs of individual patients, which is one of the drivers behind the research presented in this thesis.

Although the presented research is developed specifically for UMC Utrecht, the results are applicable in a general healthcare context, through the large network of healthcare institutes within CHOIR. This is especially shown in Chapter 5, where a generic approach is used to analyze a USA hospital in addition to UMC Utrecht.

The collaboration with UMC Utrecht Cancer Center offers the opportunity to not only take on problems that are interesting from a scientific point of view, but also to tackle challenging problems that are relevant for health care practice. The aim of the research in this thesis is to implement the proposed solutions in practice in order to add to the goal of UMC Utrecht Cancer Center to improve the quality of care for their patients. As implementation of research models and results is known to be hard in our field of study, Chapter 9 will elaborate on fruitful collaborations and key characteristics of successful research projects with impact in practice.

1.8 Thesis outline

This thesis consists of five parts, following the cancer care processes flow of Figure 1.1. Each part contains one or multiple chapters, which are introduced below.

Part I provides a general overview of multi-disciplinary appointment planning. After an introduction to the organization of processes involved in cancer diagnostics, which is given in this chapter, *Chapter 2* gives an overview of the literature on multi-disciplinary planning in health care. The literature is categorized according to the considered hierarchical level, system characteristics, variability usage, and generality and applicability. We show that many cross-relations can be identified between various healthcare applications, although no such relations are present in the current literature.

Part II consists of two chapters on the scheduling of pathology processes. In *Chapter 3* a mathematical model is developed to schedule all histopathology laboratory activities. Histopathology laboratories aim to timely provide high quality diagnostics. However, as large batching machines are in the middle of a labor intensive multiple stage process chain, the peaks in workload for employees are of concern. Therefore, we develop a decomposed MILP model, which schedules the tissue samples over the various histopathology activities, while optimizing the turnaround time and workload distribution. Using this model, a case study is performed in *Chapter 4* to show the practical applicability and to assess the improvement possibilities in UMC Utrecht's histopathology laboratory. The results show that significant improvements in turnaround time and workload division over the day can be obtained.

Part III consists of three chapters on outpatient scheduling. In *Chapter 5* we show the relation of patient no-shows and cancellations with the scheduling interval of a clinic. An extensive data analysis shows that the probability of patient cancellation and no-show increases with a larger booking horizon. Therefore, we present a queueing model to determine the optimal booking horizon, in order to reduce the effects of no-shows and cancellations. This model considers a trade-off between cancellations and no-shows on the one hand, and patient rejections on the other hand. Two case studies, in Mayo Clinic and UMC Utrecht, are provided to show the applicability of the model. In *Chapter 6* an appointment template is designed for a multi-disciplinary cancer clinic, in which the routing of patients is uncertain. As clinicians are highly valuable resources, unnecessary idle time of clinicians is undesirable. We develop a stochastic program that determines which timeslots to use for regular patients, that are known in advance, and which timeslots to reserve for multi-disciplinary patients, from which the routing is unknown in advance. The stochastic average approximation approach provides good results for a case study in UMC Utrecht's HPB clinic. *Chapter 7* continues the work of Chapter 6, and analyzes on an operational level of control how the multi-disciplinary clinic should operate. Using a computer simulation model, various planning rules are evaluated, as well as invitation strategies and patient prioritization rules. Results show that the invitation and routing strategies have the largest impact on the performance of the clinic, and that a trade-off should be made between waiting time of patients and overtime of clinicians.

Part IV consists of one chapter on treatment planning, focused on surgical treatment. In *Chapter 8* a classification scheme for surgical case mixes is given, together with a benchmark set. The case mix of a specialty or hospital influences the possible surgical scheduling performance. To visualize the case mix differences, we present a case mix classification based on the duration and coefficient of variation of the surgery types in the case mix. Furthermore, in order to allow researchers to compare their approaches and to assess whether their approach is feasible for various case mix types, a benchmark set is developed. In the instance generation process, the concept of 'instance proximity' is introduced, which allows maximizing the difference between instances.

Part V consists of two concluding chapters. In *Chapter 9* we discuss the CHOIR ecosystem and the conditions for impact of OM/OR projects in practice, based on the research presented in this thesis. The tools developed for the histopathology laboratory for example enabled the laboratory management to decide to accept the demand in laboratory work of an additional regional hospital in UMC Utrecht's laboratory. Another example is the design of the agenda blueprints for the multi-disciplinary clinics of gastro intestinal cancer in UMC Utrecht, based on the tool developed in Chapter 6. The final chapter of this thesis, *Chapter 10*, provides a conclusion to the previous chapters, reflects on the obtained results, and identifies trends in oncology operations management that may encourage future research.

CHAPTER 2

Multi-disciplinary appointment planning - a review

2.1 Introduction

Coordinating multi-disciplinary care is becoming increasingly important, especially in cancer care. As patients get more complex diseases and co-morbidities, the need for coordinated care over multiple departments increases [218]. Treatments are more and more organized as a combination of care from various disciplines or different facilities [295]. Furthermore, patients increasingly demand efficient care which is well-organized and suited to their needs. All these trends ask for an integrated approach, in which multiple disciplines together organize and optimize the patients' care pathways. This review focuses on optimization and evaluation approaches for multi-disciplinary systems.

We define a multi-disciplinary care system as a care system in which multiple interrelated appointments per patient are scheduled, where healthcare professionals from various facilities or with different skills are involved.

Cancer care is an example of multi-disciplinary organized care, as almost all cancer patients require interventions from multiple specialists, as explained in Chapter 1. Therefore, while focusing on improving and optimizing cancer care processes, research in multi-disciplinary appointment planning is of interest. This review chapter shows that planning problems in for example rehabilitation treatment and cancer diagnostics turn out to be quite similar from a mathematical point of view. In rehabilitation treatment, a patient requires appointment series with therapists from multiple disciplines, for example a physiotherapist, a psychologist, and a dietician. Furthermore, there might be precedence relations between some of the appointments, for example if physiotherapy training is required after a prosthesis has been made. Since outpatients usually have to travel far to reach the clinic and since they do not want to travel each day of the week, the challenge is to schedule as many appointments as possible on the same day, with minimal waiting time. In cancer diagnostics, this same question for a similar system is relevant, as patients require multiple consultations with various specialists in a certain predetermined order, preferably in one day, in

order to know whether there is a tumor, and if so, in what stage. In Section 2.2 we show that these so-called cross-relations are not only present in rehabilitation and cancer care, but in many healthcare settings. Since the underlying systems show similar characteristics, there is ample room for cross-fertilization between multi-disciplinary environments in healthcare.

The organization and optimization of healthcare processes got the attention from Operations Management/Operations Research (OM/OR) researchers in the past years. Especially the situation in which patients require a single appointment within a single discipline is well studied [3]. Although there are several good literature reviews on appointment planning in healthcare (e.g., [3, 31, 56, 119]), these reviews do not include multi-disciplinary planning. Vanberkel et al. [295] reviewed the literature and showed that few studies focused on multiple hospital departments. They reviewed literature on both operations research and clinical pathways, from which the first included several works on multi-disciplinary planning. Marynissen and Demeulemeester [195] reviewed the integrated systems literature, but only included hospital settings. We focus on a broad healthcare context, which for example also includes blood collection sites and nursing homes.

Multi-disciplinary planning is more challenging than single appointment planning, or multi-appointment planning for a single discipline. From a mathematical perspective there are more constraints that should be simultaneously taken into account, such as precedence relations between appointments of a variety of resources and the availability of resources from various disciplines. Furthermore, through the increasing number of resources, problems encounter a large state space and decision space. Lastly, similar to supply chain management systems, the bullwhip effect is often present in multi-disciplinary systems; Variability that occurs in early stages of a patient's care pathway, impacts the possible efficiency in later stages, something that may be relevant when scheduling multi-disciplinary systems.

The multi-disciplinary planning problem in healthcare consists of the following components:

1. Appointment characteristics: This includes the type of appointments and the resources that are required for each of these appointments. This might also include restrictions on whether patients should be treated by the same doctor or therapist during their care pathway.
2. Resource characteristics: This includes the number of resources, the discipline or skill of each resource, capacity constraints and the (non-)renewable nature of these resources.
3. Care pathway characteristics: This includes the number of patient types, the number and type of appointments required for a certain patient type and the urgency (e.g., emergency) of a patient type. Furthermore, it contains precedence constraints and time constraints that may apply to all or some of the required appointments, and states whether the appointment sequence can be changed during the treatment and if patients can recirculate

in some parts of the care pathway.
4. Objective: This includes the model objective, or set of objectives.
5. Planning characteristics: This includes the planning decision, which is either to dimension capacity, to plan capacity, or to allocate capacity to patients. This last setting also includes the decision whether appointment requests are planned immediately at arrival of the patient (online planning), or can be saved up to be scheduled once per time period (offline planning).
6. Environmental characteristics: This includes the (non) punctuality of patients and healthcare providers, the in- or exclusion of patient no-shows and cancellations, and the admission policy of patient types (e.g., is it allowed to reject patients?).

2.1.1 Focus of the review

The aim of this review is twofold. First, we provide the reader with an overview of multi-disciplinary planning and scheduling literature in the healthcare context, including the recent developments, which helps to guide further research on multi-disciplinary appointment planning and scheduling. Second, we structure the available literature based on multiple characteristics, such that researchers can easily find literature with similarities to their projects. This facilitates the comparison and cross-fertilization of approaches, as similar systems are identified.

The focus of this review is on prescriptive techniques which improve and optimize multi-disciplinary appointment systems. Prescriptive techniques include exact and approximate optimization studies, and evaluation studies, for example using simulation, which are all included in this review. We excluded descriptive and predictive approaches, which include hypothesis testing and forecasting techniques respectively.

Multiple research areas are excluded from this review. First, capacity dimensioning is not included in this review, as decisions for multi-disciplinary planning on this level are similar to these of systems where just single appointments or one discipline are involved. Multi-disciplinary capacity dimensioning involves decision making over a long planning horizon and is based on highly aggregated information. Therefore, it is not necessary to take multi-disciplinary planning constraints into account, such as constraints on resource availability, precedence constraints and interrelatedness of disciplines and appointments to make capacity dimensioning decisions. More information on capacity planning can be found in Hulshof et al. [145] or in the recent review of Ahmadi-Javid et al. [3].
Second, we do not consider personnel planning other than for capacity-to-patient assignment decisions, as personnel planning does not have different characteristics for multi-disciplinary systems than for mono-disciplinary systems. More information about personnel planning can be found in Van den Bergh et al. [32].
Third, at the capacity-to-patient level, we only consider appointment planning systems in which interrelated appointments can be planned separately. An example of a multi-appointment planning system that is not included is Condotta and Shakhlevich [74], who plan multiple chemotherapy appointments which need

to follow a specific cyclic pattern. Note that research considering the planning of chemotherapy drug injections in relation to a consult with the oncologist, and the drugs preparation in the pharmacy, is included in this review, because multiple disciplines (e.g., the pharmacy and the oncologists) are planned simultaneously.

We started our search with the review of Vanberkel et al. [295], as well as those of Ahmadi-Javid et al. [3] and Hulshof et al. [145], as these studies include multi-disciplinary appointment planning research in healthcare. Furthermore we searched the databases Web of Knowledge and Scopus for relevant papers, using combinations of relevant keywords, such as *appointment planning*, *scheduling*, *multi-disciplinary*, *one-stop-shop*, *rapid diagnostics*, *calendar planning*, *flow shop*, *open shop*, and *flexible shop*. For any article found, we performed a forward and backward search to find additional manuscripts. We limit the review to papers that are written in English and are published before January 1st, 2017. The search procedure resulted in a set of 63 articles, which are all classified in the Orchestra database (www.choir-ut.nl).

2.1.2 Structure of the review

To identify cross-relations, we start this survey with a description of the healthcare applications in multidisciplinary planning in Section 2.2. Following Beliën and Forcé [27] and Cardoen et al. [51], the remainder of this literature review is based on different perspectives to analyze all included articles. In this way, a researcher can query a list of papers according to specific needs and interests. These so-called classification fields are descriptive, and include problem characteristics, solution characteristics, and system characteristics. Each section discusses one classification field, together with (a selection of) all relevant manuscripts. A manuscript is therefore discussed from various perspectives [51], and researchers can focus on the classification field of their interest [27].
We consider the following classification fields:

1. Decision delineation/hierarchical level (Section 2.3): reviewing the literature based on various planning decisions, at various hierarchical levels.
2. System characteristics (Section 2.4): reviewing the literature based on precedence constraints included in the problem context (flow-shop, open-shop, and mixed-shop systems).
3. Variability (Section 2.5): reviewing the literature based on the incorporation of uncertainty and variability.
4. Generality and applicability (Section 2.6): reviewing the literature based on the scientific impact (benchmarking) and the impact in practice (case studies, implementation).

Each section starts with a short description of the classification field and the distinct areas on which manuscripts are differentiated. Furthermore, the relevant literature in each of these areas is discussed, and a table is provided to categorize manuscripts in this classification field. This review ends with Section 2.7, which provides a conclusion and open research challenges.

2.2 Healthcare applications

An integrated view is essential for optimizing the care chain from a patient and provider perspective. In the literature, we see that multi-disciplinary planning is increasingly introduced in healthcare settings. In Section 2.2.1 we explore the motivation behind the implementation of multi-disciplinary care, as multi-disciplinary systems are well represented in the medical literature. We identify several application areas and cross-relations in Section 2.2.2. We conclude with directions for further research in Section 2.2.3.

2.2.1 Motivation for organizing multi-disciplinary care

There are several reasons for healthcare systems to introduce multi-disciplinary care in their systems. The first and most heard argument is to provide patient centered care. Therefore, hospitals focus on improvements in patient satisfaction and quality of care [182]. Patient satisfaction is quantitatively measured in terms of access time [113], and waiting and throughput times [14, 108, 283]. A general pattern is observed that most multi-disciplinary systems in the medical literature are focused towards providing all consultations on a single day. Quality of care is for example measured in number of changes in prescriptions or diagnoses, and adverse outcomes [99, 107], as more coordination between clinicians is believed to result in fewer mistakes and more first-time-right diagnoses [99, 108].

The second reason for healthcare systems to introduce multi-disciplinary care, is the structure it provides to the system. The implementation of multi-disciplinary care is a means to force coordination between various healthcare units, and enables to focus on a specific group of patients [211].

To facilitate structure in healthcare settings, easily adoptable tools are preferred. Therefore, researchers should include this requirement in the design of multi-disciplinary planning tools, such as planning software or decision rules. Simple planning solutions are most often the easiest to implement and understand for the healthcare staff that has to work with the tools. This way, structure and coordination can be provided, together with an increased planning efficiency.

A third reason to introduce multi-disciplinary care, is to facilitate a new clinical practice, which involves clinicians from multiple specialties, such as an intake for ambulatory Huntington's disease patients [302], the follow-up for children with neuromuscular diseases [303], or a multi-disciplinary cancer clinic (see Chapter 6 and 7). Under these circumstances, it is hard to compare the performance of the new system design against practice, as the initial performance does not reflect the performance of the new system. Researchers are therefore challenged to show that their design will perform well in practice, compared to other reasonable design options.

2.2.2 Application areas

Multi-disciplinary systems are present in a variety of healthcare settings. In this section we show that multi-disciplinary care knows many applications, and that the organization of this care, and more specifically the relevant underlying characteristics, show similarities. We found the following application areas:

1. Outpatient and day care clinics
2. Cancer clinics
3. Rehabilitation clinics
4. Emergency patient care
5. Elective patient care
6. Care processes without a patient present
7. Blood collection sites

Outpatient and day care clinics provide non-overnight care for patients. A trend recently introduced in these clinics is to organize care in a patient centered way. This can facilitate personalized diagnostics and treatment, and increases patient satisfaction. The concept of a flow-shop, where multiple consecutive consultations are offered, is therefore often seen in outpatient and day care clinics, especially for patients with regular checkups, or when patients need an intake or diagnostics [107, 301, 302]. [107] describes an epilepsy transition outpatient clinic, where staff from multiple disciplines consult patients. Not only single provider consultations, but also consultations with multiple providers at the same time are offered. The clinic operates as a flow-shop, in which all consultations are consecutively scheduled, such that the waiting time for patients is minimized, followed by a diagnostic work-up if needed. [301] and [302] designed an outpatient clinic to facilitate patients with Huntington's disease with an individual treatment plan. Here, all relevant care providers will see a patient in a predefined order during a visit to the outpatient department. A chemotherapy day care clinic also requires involvement of multiple departments in the treatment of patients. As the planning of the drug preparation by the pharmacy and the drug injection by the nurses should be well-coordinated, an integrated perspective is required in planning the chemotherapy appointments [169]. Other examples of multi-disciplinary outpatient and day care clinics can be found in cancer diagnostics [108], neurology [113], nuclear medicine [236] and ophthalmology [181].

A patient in a *cancer clinic* needs a diagnosis, personalized treatment plan, and treatment. As many specialties are involved in the diagnostic trajectory of a cancer patient, the treatment opportunities are discussed with a variety of disciplines during a multi-disciplinary meeting. Nowadays, hospitals realize that not only the treatment plan should be developed by a multi-disciplinary team, but also that the patients want to meet this team, and receive all relevant information for their treatment from this team [182]. Therefore, multi-disciplinary clinics are designed, in which a patient can meet with any relevant clinician for their treatment, as well as with other providers such as psychologists, dieticians, and social workers if needed. The challenge in the organization of these clinics is

that patients only need to consult a subset of clinicians from a multi-disciplinary clinician pool, whereby this subset is known at a very late moment in time and should get a consultation within a small time frame [177].

In a *rehabilitation clinic*, patients with various movement disorders are treated. The rehabilitation treatment consists of appointment series with therapists from various disciplines during several weeks or months, coordinated by a rehabilitation physician. Once every several weeks, the physician and all involved therapists discuss the progress and possible adjustments in the treatment. Scheduling the appointments is challenging, as patients prefer to combine several treatments on one day, while they have fixed therapists for every discipline. In the organization of these treatment pathways challenges are, amongst others, the continuity of the care process, a simultaneous start for all disciplines and a short access time [40].

Emergency patient care considers patients that need (semi-)acute care. To triage and diagnose these patients, they often need multiple tests, which can be performed in various orders, represented by an open-shop or mixed-shop system. Multi-disciplinary planning is involved on an online decision level, not only with respect to the timing of the tests, but also to the sequence of the tests [16].

Elective patient care considers patients that need a planned intervention, such as surgery. Multi-disciplinary planning is done at several levels for this patient population. First, the relation between the outpatient clinics, the operating room, and the wards is relevant. Capacity shortage in one area, may lead to waiting lists or emptiness in other areas. Second, inpatient care services for hospitalized patients require efficient planning when diagnostic tests and treatments are required from multiple departments [75]. In this case, it is important to minimize a patient's length of stay, as each occupied bed blocks the access to care for another patient. Finally, multi-disciplinary planning can be approached from an opposite direction. Instead of a patient that has to visit multiple types of providers, a provider has to visit multiple types of patients. For example in patient-to-nurse scheduling at the wards, which can be represented by an open-shop system, time constraints are restricting the possible schedules [66].

In most *care processes without a patient present*, such as laboratories and sterilization plants, patients are processed in a fixed activity sequence, where various resources are required for the activities [175, 263]. Applications from the laboratory, and, on a higher level from the process industry underlying the laboratory process optimization research, can be used in optimizing outpatient clinics. However, the difference between an outpatient clinic and a laboratory is the level of variability on the capacity-to-patient assignment level. Where laboratories are highly automated, and therefore have activities that are well predictable, patient consultations are provided by people. Therefore, laboratories experience less variability in the activity duration.

Blood collection sites are flow-shop type systems with even more variability, as not only variability in activity duration, but also variability in donor arrival has to be taken into account. In line with the laboratory, blood collection from donors requires a fixed series of activities. These activities are often performed by

the same staff, but in some countries, such as France, multiple different providers are required since the various activities have to be carried out by certified staff members. In these cases, the design of a blood collection system requires a multi-disciplinary appointment planning approach. As blood donations are often voluntary, high service levels are required to ensure satisfied donors. Therefore, the donor flow through the system needs to be well designed, and matched with the staffing requirements [5].

Cross-relations between the various application areas are rarely reported upon. However, five manuscripts are presented in a generic way, without one specific application area mentioned. [309] analyze the patient flow through a hospital, which is applicable to the emergency and elective patient flow. [306] and [313] consider the scheduling of multiple appointments for multiple patients of various patient types on the same day, a problem which is relevant to the rehabilitation clinics, cancer clinics, and for example ward scheduling. [25] and [146] consider an elective patient admission problem with multiple resource requirements and constraints. This is for example applicable to outpatient clinics, cancer clinics, and the planning of the elective patient care chain. [25] present a case study of a neurosurgery department, to show the applicability of their method, whereas [146] apply their approach to generated data, representing many different health care settings.

2.2.3 Conclusions and further research

Multi-disciplinary care systems are present throughout the hospital, from outpatient clinics to laboratories. They are introduced for several reasons, including improved patient centeredness, improved structure and coordination, and to facilitate clinical improvements.

Despite the different application areas, design and optimization insights can be gained by comparing the underlying planning decisions in these areas. However, crossovers are rarely reported upon, as until now new methods are frequently developed for one specific application area. This offers many opportunities for further research, as a general method that can be applied to several application areas with good performance is of great value to healthcare professionals.

As an example, insights from the research on the planning of outpatient clinics and cancer clinics with variable resource requirements, such as clinics where patients may need immediate extra tests depending on the results of previous testing [177], are also relevant for treatment planning, for example in a rehabilitation setting. Both application areas can benefit from research into the question on how to deal with an unknown patient pathway and unknown need of resources.

A second example is the question on how to minimize the length of stay for patients. This question is relevant for inpatient care planning, by planning several diagnostic tests and treatments over a couple of days. This question is also relevant for rapid diagnostic trajectories, where cancer patients need to be provided with a diagnosis as fast as possible. Both these areas could therefore benefit from each other, via cross-relations and shared research results.

Most reported application areas are located within a hospital. Multi-disciplinary healthcare areas outside hospitals are interesting areas for further research. Examples are blood and transplant management, transmural care, home care, and nursing homes. Again, these application areas have similar questions and a similar structure as known multi-disciplinary systems. Blood collection sites for example share commonalities with laboratories and outpatient clinics, and nurses in a home care environment need the same type of planning as nurses in wards.

2.3 Hierarchical level

Multi-disciplinary planning can be considered at different hierarchical levels:

1. Capacity dimensioning (long-term)
2. Capacity planning (mid-term)
3. Capacity-to-patient assignment (short-term)

Capacity dimensioning involves decision making over a long planning horizon and is based on highly aggregated information. As described in Section 2.1, capacity dimensioning is not included in this review, since decisions on this level are similar to mono-disciplinary systems. More information and articles on capacity dimensioning decisions can be found in Hulshof et al. [145] or in the recent review of Ahmadi-Javid et al. [3].

Capacity planning specifies the results of capacity dimensioning decisions into a division of the resource capacity to patient groups or time slots [127]. In this way, blueprints for the capacity-to-patient assignment are created in which resources are allocated to different tasks, specialties and patient groups. Patient admission policies and temporary capacity expansions such as using overtime or hiring staff are also part of capacity planning.

Capacity-to-patient assignment involves the appointment planning at the individual patient level [145]. Following the blueprints, a date, time, and resources are allocated to a specific patient.

Note that the decision horizon lengths are not explicitly given for any of the planning levels, since these depend on the specific characteristics of the application. For example, in a one-stop-shop diagnostic setting, horizons will be shorter than in rehabilitation care where treatment takes several months.

We found 19 papers on capacity planning, which are described in Section 2.3.1. Furthermore, we found 49 papers on capacity-to-patient assignment, as described in Section 2.3.2. Section 2.3.3 concludes and provides opportunities for further research. Table 2.1 gives an overview of the papers and categories.

2.3.1 Capacity planning

Capacity planning considers the division of resource capacity to specialties, patient groups or time slots. This can be done by several means:

Table 2.1 Hierarchical level

Hierarchical level	Focus	References
Capacity planning	Blueprint schedule	[34, 88, 172, 175, 177, 178, 218, 222, 223, 245, 250, 315, 328]
	Patient admission planning	[25, 75, 146, 147, 262, 303]
	Temporary capacity changes	[146]
Capacity-to-patient assignment	Offline scheduling	[16, 55, 66–69, 75, 76, 92, 103, 106, 143, 148, 175, 181, 197, 198, 236, 239, 241, 254, 263, 265, 266, 269, 273, 277, 303, 313, 325]
	Online scheduling	[15, 17, 25, 40, 50, 82, 88, 125, 153, 156, 178, 196, 222, 223, 232, 235, 236, 305, 306, 320]

1. Blueprint schedule
2. Patient admission planning
3. Temporary capacity changes

A *blueprint schedule* describes the amount of capacity on a day or particular time slots that can be used for specific patient types in the operational planning. It can also be used to plan combination appointments, which are appointments where more than one doctor or therapist should be present. *Patient admission planning* considers the design of an admission policy that describes how many and which patients should be admitted from the waiting list. Developing guidelines for *temporary capacity changes* in case of demand peaks and drops is also considered as capacity planning [145].

Designing a *blueprint schedule* as a guideline for appointment planning is done with objectives to combine consultations on one day [88], to minimize waiting time on a day [178], or to minimize access time or throughput time [34, 175]. In the blueprint, time slots are assigned to patient types [34, 88, 178], or to process stages [175]. Furthermore, the blueprint may prescribe when doctors can best have consultation hours [34, 178]. A blueprint is usually designed based on expected arrival patterns and expected availability of capacity. Robustness to different patient arrival realizations is considered an important characteristic of blueprints [177]. Suitable methods to design blueprints are mathematical programming or heuristics, in combination with robust optimization or computer simulation to ensure robustness. Also stochastic programming can be used, which

takes robustness to several scenarios into account. [88] creates blueprints for a veteran clinic, where patients have to travel far to see a doctor or dentist and therefore prefer to combine several consultations on a day. In the blueprint, slots are kept open in order to plan such combinations of treatments. [178] creates blueprints for the scheduling of patients who need an appointment with an oncologist followed by chemotherapy treatment. A radiotherapists' schedule for consultations and scan reviews of different patient types are designed in [34], to ensure timely treatment for all patient types. Blueprints that prescribe the order of tasks to be performed in a laboratory setting are designed in [175].

Used methods include mathematical programming [34, 175, 178, 250] and heuristics [88, 175], simulation [34, 88, 172, 178, 245, 315], queueing theory [315] and stochastic programming [250].

In some multi-disciplinary systems, patients require one or more combination appointments, that is, a single appointment where more than one doctor or therapist should be present. Examples are group therapy in rehabilitation, where a group of patients is treated by multiple therapists from the same or different disciplines, and multi-disciplinary team meetings, where the diagnosis or treatment is discussed with or without the patient's presence. For these systems, it is essential to align staff schedules, for example by means of a blueprint schedule, to ensure that members of a multi-disciplinary team have enough options for combined care or meetings. We found only one article where this problem is addressed: [218] align staff schedules in surgical cancer care, using mathematical programming.

Patient admission planning considers the design of an admission policy. The treatment of skin cancer is considered in [262], where patients are admitted immediately to either a regular consultation or a one-stop-shop consultation, depending on their medical characteristics and the already booked capacity. [25] base the admission of elective patients for surgery on expected profit. If (semi-)urgent patients may still arrive after the patient admission decision, it might be worthwhile to take future scenarios into account in admission planning [147].

Used methods include mathematical programming [75, 146, 303], simulation [262], queueing theory [303] and dynamic programming solved with heuristics [25].

Temporary resource capacity changes are increases or decreases in capacity allocation during a specific time frame, to cope with fluctuations in patient demand [186]. Temporary capacity change can improve the balance between access times and resource utilization [307]. This topic is not widely studied in healthcare, but it is important for a good healthcare planning and control [145]. Especially in the multi-disciplinary case, such a balance is essential to avoid large bullwhip effects in related disciplines. We found one paper studying temporary capacity changes in a multi-disciplinary setting. [146] design allocation policies for resources that divide their time over multiple tasks in the care chain, based on the patient's waiting list status and access time target. The used method is linear programming [146].

2.3.2 Capacity-to-patient assignment

For capacity-to-patient assignment, we distinguish *offline planning*, where planning requests are saved up and executed once per period, and *online or advance planning*, where an immediate response is required to each current incoming request. The decision of planning offline or online is a management choice, where trade-offs have to be made between high utilization (mostly achieved in offline planning) and short response times (mostly achieved in online planning). Online systems are more common in practice, while the offline approach has received more attention in the literature as it is easier to model [3]. Applications of both planning methods can be found in all applications of multi-disciplinary care, but online planning is mostly reported in emergency care and cancer diagnostics and treatment, for a quick response is essential [41].

Planning decisions either focus on determining time slots for appointment series that take place on one day, or on determining both days and time slots for appointment series at the same time.

We found 33 papers on *offline planning*. Most articles focus on scheduling appointments on one day and in particular for outpatient departments [67, 143, 181, 196–198, 241].

Scheduling series of up to twenty diagnostic and treatment appointments on one day with no specific order is done for an oncology center [196, 197] and several diagnostic facility outpatient clinics [67, 103, 143, 198], minimizing both patient and doctor waiting time. A block appointment system is used in [181] to schedule treatment appointment series in a specialty clinic, where patients are assigned to arrive at the start of a time block. Scheduling tasks on one day is also done for laboratories [16, 277] and sterilization practices [263].

Scheduling days and time slots for appointments in a several day care path applies specifically to rehabilitation care and inpatient care. Multiple studies schedule rehabilitation appointments on an inpatient and/or outpatient basis [68, 69, 254, 273]. Often, a multi-stage model is used to reduce the problem complexity. [273] formulate a three-stage model where patients are accepted or rejected for the treatment, after which therapists and time slots are determined. [254] create treatment schedules on a week level, to be specified later in terms of morning/afternoon appointments and time slots. Scheduling a series of procedures in care chains for inpatients, such as diagnostic activities and surgery, in a several week horizon is done in order to maximize the contribution margin [106], to minimize the length of stay [265] or in a day horizon to minimize waiting times and overtime [148, 313]. Series of examinations for vascular checkups are scheduled in [76], either on one day or on multiple days, for inpatients as well as outpatients. [66] schedule various (partially ordered) nurse tasks that have to be performed in a day horizon.

Offline scheduling problems are often NP-hard or NP-complete, which makes them difficult to solve exactly. Therefore, most authors use heuristics and/or decomposition into hierarchical subproblems to solve the problem.

Used methods include mathematical programming [16, 67, 69, 76, 106, 148,

181, 197, 236, 254, 265, 273, 313], heuristics [67, 76, 143, 148, 181, 198, 241, 277, 313], simulation [15], genetic algorithms [16, 68, 69, 148, 197, 241], local search methods [66], and data mining [15, 68].

For *online* planning, we found 21 papers. Planning appointments on one day applies for example to cancer clinics and emergency departments laboratories. As an example, [153] consider the scheduling of appointments where examinations take place in the morning, after which diagnoses and treatment plans are determined in multi-disciplinary team meetings and the outcome is discussed with the patient in the afternoon. Rescheduling is allowed in [305], where involved departments may change a concept schedule, and in [17] where an arriving patient in a pathology emergency department laboratory is scheduled and other patients are rescheduled, such that the total waiting time of all patients is minimized.

Online planning of appointments on multiple days is done for rehabilitation and cancer treatment. [40] present a methodology to plan appointment series for rehabilitation outpatients. The scheduling of the radiotherapy care pathway is considered in [156] and [320], where operating hours of treatment machines and shift hours of machine operators are varied [156] or different arrival distributions and oncologist productivity are considered [320]. [125] use a template for scheduling chemotherapy patients online, and if the request does not fit in the template, it is updated.

For practical applications of online planning, a short computation time is essential since an immediate response is given to the patient. Therefore, several papers use heuristics for optimization or simulation for evaluation, in order to obtain a reasonable good solution with a short computation time.

Used methods include mathematical programming [17, 40], heuristics [82, 235], simulation [82, 153, 156, 235, 320], stochastic programming [236], constraint programing [125], genetic algorithms [17], and agent-based models [305, 306].

2.3.3 Conclusions and further research

Mid-term capacity planning in healthcare has received relatively little attention compared to capacity-to-patient assignment [3], and the same holds in multi-disciplinary care. However, mid-term capacity planning is essential for a good healthcare system control [127].

Three topics in particular are a promising direction for further research. The first is the alignment of staff schedules, for which we only found one paper. However, many applications exist where patients require one or more combination appointments, such as group therapy in rehabilitation, and multi-disciplinary team meetings in diagnostics or treatment planning. For these systems, it is essential to align staff schedules, for example by means of a blueprint schedule, to ensure that members of a multi-disciplinary team have enough options to deliver combined care or to attend joint meetings.

A second direction for further research are temporary capacity changes. Temporary capacity changes are not widely studied in healthcare, but are important for a good healthcare planning and control [145], as they can restore the

balance between access times and resource utilization [307]. Especially in the multi-disciplinary case, such a balance is essential to avoid large bullwhip effects in related disciplines. Since the problem of when to change capacity and to which extent involves optimization over time, dynamic programming could be a suitable method for the optimization of temporary capacity changes. As multi-disciplinary care systems involve complex state descriptions and many possible actions, this should be combined with approximation methods.

The last direction for further research considers online planning in capacity-to-patient assignment. We observe that online planning is frequently studied in a simulation setting, where several scenarios are evaluated. However, only few optimization studies are known. This would be a promising topic for further research, since many applications exist for online planning. Taking future patient arrivals into account in optimization can be done by dynamic programming, stochastic programming, or robust optimization. Using these methods, a combination with approximation methods would be required since the planning of multiple appointments involves many possible actions.

2.4 Type of system

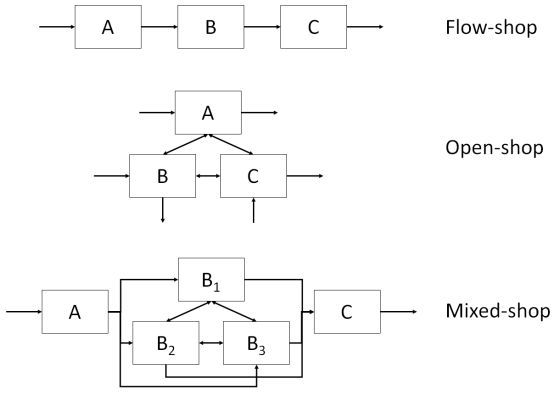
Precedence relations and time constraints between appointments may be present in multi-disciplinary appointment planning. For example, a patient first has to finish all diagnostic tests, before a consultation with a specialist is planned. We distinguish three different multi-disciplinary systems, based on precedence constraints, since each of these systems faces different optimization problems:

1. Flow-shop
2. Open-shop
3. Mixed-shop

In a *flow-shop system*, also referred to as one-stop-shop or carousel, patients undergo a predefined sequence of activities at multiple facilities. Through the fixed sequence of appointments, service process divergence is low, as there is a high degree of standardization. Furthermore, there are strict precedence relations between activities. An example of a flow-shop can be found in a specialty clinic for Huntington's disease, where multiple symptoms need to be addressed in consultations with several professionals, scheduled in a predefined sequence [301]. Also in one-stop-shops for cancer diagnostics, flow-shop systems are often present, as patients follow a predefined trajectory.

In an *open-shop system*, patients undergo a set of activities which can be scheduled in any order. Through the flexibility in order of activities, service process divergence is high, as each patient can get the appointments in a different order. Open-shop systems contain few or even no precedence constraints. An example of an open-shop can be found in rehabilitation treatment, where patients need an appointment with a physician, physiotherapist, and psychologist, in an arbitrary sequence, as long as these appointments are planned on the same day

Figure 2.1 Visualization of a flow-shop, open-shop, and mixed-shop system



[117] or in the same week [40].

A *mixed-shop system* is a combination of a flow-shop and an open-shop system, with an intermediate level of service process divergence and prevalence of precedence relations. A mixed-shop regularly has a fixed sequence of "consultation - examinations - consultation", but the order in which the examinations take place is variable. An example is a regular diagnostic trajectory, where a patient first has an intake consultation, thereafter has multiple examinations in an arbitrary order, and finishes with another consultation in which the results are discussed [197, 305].

Figure 2.1 presents a visualization of the three system types. In each of these three systems, patients can undergo the complete set of activities, or only a subset of activities that are applicable for them. We call this subcategory a flexible shop. For example in a flexible flow-shop, patients can skip an examination that is not applicable for them, and continue with the next one. This way, patients can undergo a different subset of activities, albeit in the same sequence, as the flow-shop system is still present. In a flexible open-shop, patients can also undergo different subsets of the activities, in any order. These flexible shop systems are especially relevant for personalized healthcare settings.

Note that in each of the aforementioned systems, the patient-to-doctor approach, in which a patient visits multiple clinicians in a row, is most prevalent in the literature, although the doctor-to-patient approach, in which multiple clinicians visit a patient, could be applied as well.

Out of the total of 63 papers found, 36 papers consider a flow-shop system, as described in Section 2.4.1, 9 papers consider an open-shop system, as described in Section 2.4.2, and 18 papers consider the mixed-shop system, as described in Section 2.4.3. Each section gives an overview of the literature, and provides the patient and system measures that are of importance when studying the discussed system. Section 2.4.4 provides conclusions and opportunities for further research. Table 2.2 gives an overview of the papers and categories.

Table 2.2 Type of system

Type of system	References
Flow-shop	[15, 17, 25, 34, 40, 50, 55, 82, 92, 106, 146, 147, 153, 156, 175, 177, 178, 181, 218, 222, 223, 235, 236, 239, 245, 262, 263, 266, 277, 304, 309, 313, 315, 320, 325, 328]
Open-shop	[16, 88, 143, 148, 197, 198, 232, 250, 305]
Mixed-shop	[66–69, 75, 76, 103, 125, 155, 172, 196, 241, 254, 265, 269, 273, 303, 306]

2.4.1 Flow-shop

Flow-shop planning includes the planning of one-stop-shop, rapid diagnostics, and carousel programs, in which predefined care pathways are present. These care pathways most often span a single day, but in specific healthcare areas, such as for patients with head and neck cancer, rapid diagnostics programs of multiple days are designed [283].

The large amount of flow-shop papers considers a deterministic variant of the problem on the capacity-to-patient level. Two solution strategies are frequently applied: ILP optimization evaluated by discrete event simulation, and heuristic approaches.

Patient performance in flow-shop systems is measured by means of the direct waiting time [17, 153, 178, 181, 222, 223], access time (also known as indirect waiting time) [34, 50, 147], and the rejection probability [5]. In single-appointment systems the direct waiting time is measured as the waiting time on the day of the appointment from the planned appointment start until the actual start of the appointment. However, in a multi-appointment setting, the waiting time until the start of the first appointment does not cover all waiting on that day. Therefore, we define direct waiting time as the waiting time spent in the waiting room starting from the scheduled start time of the first appointment until the moment that the patient leaves the hospital. Therefore, it not only includes waiting time caused by a late appointment start, but it also includes the waiting time between two subsequent appointments. Besides direct waiting time, [181] also include the number of patients in the waiting room in their objective, as a measure of patient satisfaction. As all appointments in a flow-shop are scheduled sequentially, minimizing the throughput of a system, as for example studied by [262], has similar results.

Accessibility, represented by the access time is a second objective in flow-shop systems [34]. As many health systems have access time requirements for outpatient and treatment clinics, in order to ensure that patients are seen on time, access time measures become increasingly important.

If allowed for by the system, the rejection probability is included as a performance metric, as a rejection has negative consequences for the patient in terms of quality of care and patient experiences. For example blood collection systems may need to reject donors when blood inventory levels are high enough. However,

this comes at a cost of losing a donor who could be needed at a later moment in time [5].

System performance in flow-shop systems, is measured by means of completion times, throughput times, tardiness, the number of patients admitted, utilization, and overtime [25, 50, 147, 175, 178, 263, 266, 277]. Due to the sequencing relations in a flow-shop system, flexibility in the planning is limited. This influences the timeliness of care. Therefore, time-related objectives, such as completion times, throughput times, and tardiness are a set of frequently studied objectives. These time-related objectives are especially relevant in healthcare settings where patients are not physically present, such as laboratories and sterilization departments [175, 263]. When patients are physically present, the number of patients admitted, utilization, and overtime are relevant measures [25, 50, 178, 266]. When a fixed capacity is reserved for such a one-stop-shop, rapid diagnostics, or carousel program, the maximum number of patients to be admitted is restricted [262]. An important objective in flow-shop planning is therefore to maximize the possible number of patients treated [40, 146, 313].

2.4.2 Open-shop

Open-shop planning includes the planning of multiple examinations or consultations, in which the order of these appointments is not relevant. Multiple resources are required for the appointments, hence coordination between resources is required. Furthermore, in many situations, patients with various characteristics use the open-shop system, which makes the flexible open-shop system most prevalent both in practice and in the literature. For general information and applications of open-shop systems, refer to [12].

Most open-shop papers consider a flexible variant of the problem on the capacity-to-patient level. The most frequently applied solution strategies are (local search) heuristics.

Patient performance in an open-shop is measured by the number of same day appointments, an equal spread of appointments over the days, and the timeliness of care. Patients who visit a healthcare institute for multiple appointments with different healthcare providers, prefer their appointments in the same day [88]. Therefore, the number of same day appointments, is a relevant optimization criterion. However, in some situations, it is impossible to provide all needed care in one day. In this case, a care pathway of multiple days, with an equal spread of the number of appointments over these days may be more desirable, to level the care load. Where flow-shop planning gives patients more clarity about their care pathway, the benefit of open-shop planning over flow-shop planning is the possible gain in waiting time. Therefore, the timeliness of care is an important objective from a patient perspective for open-shop systems, which is shown in objectives such as access time, throughput time, and direct waiting time [16, 143, 148, 232].

System performance in an open-shop is measured by completion times [16, 305] and makespan [197, 198, 232]. By minimizing completion times and makespan, hospitals aim to minimize waiting times for the patients, and maxi-

mizing efficiency for the hospital. These indicators are especially relevant in an open-shop, as the completion time and makespan of a patient is influenced by the sequence of the needed activities [197].

2.4.3 Mixed-shop

Mixed-shop systems include appointments with precedence constraints, but with some flexibility in the sequence of a subset of all appointments. It can be subdivided in two general situations. First, we have a diagnostic facility, in which each patient first requires an intake consultation, then multiple tests in an undefined order, and finishes with a consultation again in which the diagnosis is explained [197]. Second, we have a specialty clinic in which each patient type requires specific treatments to be given in a specific order, together with some general treatment modalities for which the order is irrelevant and which can be scheduled at any free moment during the treatment period [69].

Most mixed-shop papers focus on the capacity-to-patient level, and behave as a flexible mixed-shop system. Mathematical programming is a frequently applied solution technique, as well as heuristics.

Patient performance is frequently assessed by the direct waiting time [68, 69, 269], length of stay [75, 76, 103], and leveled care load [254]. A focus on direct waiting times ensures quick access for patients on the day of the appointments. An interesting approach is adopted by [66], who assess the timeliness of care by taking both the tardiness and earliness into account. Notably, no authors have evaluated the accessibility in mixed-shop systems from an access time perspective. The patient's length of stay is minimized in order to reduces all unnecessary delays for the patient [76]. The length of stay as a performance measure is relevant in one-day diagnostic trajectories [76], and inpatient clinics, where hospitalized patients need inpatient care services spread over a few days with night stays [75]. In this case, it is important to minimize a patient's length of stay, as each occupied bed blocks the access to care for other patients. The spread of treatment appointments over multiple days to level the care load for patients is specifically prevalent in mixed-shop systems [254], in contrast to the objective to plan as many appointments as possible on one day, which is seen in open-shop systems. This is especially relevant in treatment situations that take multiple days, weeks, or even months, such as in rehabilitation care.

System performance in mixed-shop systems is evaluated by the number of patients admitted to the system [75, 254, 273, 303, 306], makespan [69, 76, 265], and completion times [196, 269]. The number of patients admitted to the system is specifically seen in specialty clinics, such as rehabilitation care [254, 273]. Minimizing the makespan or completion times is frequently studied when analyzing outpatient facilities [69, 196, 265, 269]. These measures are chosen to optimize the operational efficiency.

2.4.4 Conclusions and further research

Despite the high prevalence of flow-shop systems in the multi-disciplinary literature, flexible flow-shop systems are not reported upon. As seen in Section 2.2, the organization of patient centered clinics and personalized diagnostics and treatment systems is a gap in the literature that deserves considerable research attention in the near future. These systems ask for more flexibility, by only selecting the required steps in a flow-shop system that fit the needs of the patient.

Open-shop planning in healthcare has not received much attention in the literature so far. It requires high flexibility and coordination of all participating resources, without them being able to fix capacity for specific patient groups. However, the joint optimization of multiple disciplines rapidly attracts more attention of researchers and of practice, with large improvement possibilities. The implementation of structured pathways, such as flow-shop systems has received much attention in the medical literature. This comes at a cost of reserving capacity for specific patient groups [176, 297]. Therefore, we expect open-shop systems to become of more interest for researchers and practitioners in the near future. As healthcare systems involve more complex behavior than those open-shop systems that are polynomial solvable, research in approximation methods and intelligent optimization techniques is promising [19]. To ensure implementation, individual resource performance should be analyzed as well, to show the individual disciplines the benefits or costs of coordinated care for themselves.

Few papers studied the combination of access time and direct waiting time or throughput time [34, 40, 303]. This offers interesting opportunities for further research. Flow-shop systems, such as the one-stop-shop, are most often designed for specific days of the week, in which the direct waiting time might be low, but access times might take up to a week. On the other hand, open-shop systems might offer direct access to the first activity, but may end up with a long throughput time. It is an open challenge to develop optimization methods that ensure a good fit between the access times and direct waiting times.

We have seen that in flow-shop systems and mixed-shop systems, the number of patients that gets access to the system is a relevant performance indicator. Access times in mixed-shop systems are especially an interesting area of future research, as no such literature is known at this moment. For flexible shop systems, accessibility as an indicator is more difficult to include. This would create an unfair access policy over different patient types, as it is advantageous to grant short-stay patients access more frequently than long-stay patients.

When a care pathway consists of multiple days, other performance indicators become of interest compared to one-day care pathways. For example an equal spread of the number of appointments over these days is desirable. Until now, this has not been subject of research, and offers opportunities for further research by adapting used methods to use this new objective.

Similar to mono-disciplinary appointment planning, most multi-disciplinary studies consider an objective function with multiple criteria, involving patient as well as system performance. Most studies sum weighted objectives to derive

the final objective value. Other multi-objective optimization methods, for example with non-linear relations between performance indicators, are still an open challenge in multi-disciplinary appointment planning.

It is well known that hospitals tend to evaluate performance on a local, departmental, level. Therefore, in order to ensure implementation of all three type of systems, not only the entire system's performance, but also the individual resource performance is of interest. Resource idle time, overtime, and utilization are thus relevant performance indicators, although not yet considered in the literature. For example, outliers in individual resource performance might indicate that changes in capacity or opening hours are required.

2.5 Variability and uncertainty

In modeling a planning problem, researchers have to decide whether to take variability into account or not. This decision depends on both the extent to which variability is a characteristic aspect of the planning problem, as well as the model complexity. Variability in multi-disciplinary care exists in various aspects of the system:

1. Patient arrivals
2. Appointment durations
3. Resource capacity
4. Care pathway

For *patient arrivals*, both the number of arriving patients and their moment of arrival are subject to variation. If patients are assigned a series of appointments, no shows and late arrivals might occur. For walk-in patients, certain time slots might be more popular than others. When scheduling an appointment, the *appointment duration* per patient or appointment type is often considered fixed. However, in practice these durations can vary to a greater or lesser extent. *Resource capacity* can vary due to longer term reasons such as sabbaticals or parental leave, or shorter term events such as illness or machine breakdowns. A patient's *care pathway* is either known at the moment of arrival (e.g., patients need predetermined appointments for treatment and yearly check-ups), becomes clear during the appointment series (e.g., when a diagnosis is involved) or is modified along the way (e.g., the course of a rehabilitation treatment might depend on the patient's progress, or a patient might recirculate in some parts of the care pathway).

The extent in which variability and uncertainty are prevalent in the described aspects depends among others on the planning decision (see Section 2.3) and the system type (see Section 2.4). Information with respect to arrivals and care pathways of individual patients is usually not yet available for long-term and mid-term decision making [129]. The same holds for information about the (detailed) availability of resource capacity. In order to make decisions, this information has to be forecasted. In short-term decision making, such as capacity-to-patient assign-

ment, more information is available, since (part of) the patients already arrived and staff schedules are made [129]. All information with respect to arrivals, capacity and care pathways for a time period is gathered in online planning, although information with respect to future arrivals is still uncertain. Therefore, variability in arrivals, care pathways and available capacity are characteristic aspects of problems considering capacity planning and capacity-to-patient assignment in an online setting.

Variability in appointment durations is a characteristic aspect in problems with time constraints or objectives concerning waiting time, idle time or overtime. In these problems, not taking variability into account may influence the robustness of the obtained solution in practical situations.

The system type impacts the incorporation of variable aspects as well. In a flow-shop system, the order of appointments is fixed. However, in an open-shop system, the order of appointments is unknown until the appointments have been scheduled. This allows for (re-)scheduling on the day of the appointments. For example, by deciding upon the order of the appointments according to realized durations of previous appointments and current waiting times.

Researchers can model variability in different ways. A deterministic approach assumes that all information pertaining to the variable factors is known with certainty at the time of the decision making. A stochastic approach considers uncertainty in these variable factors. Adopting the stochastic approach in modeling often significantly increases a model's complexity, which may result in an 'exploding state space' and large computation times. The solving time is often important for practice, especially for capacity-to-patient planning and in particular online planning [41].

In the next sections, we give an overview of the extent to which the literature takes variability into account in each of the described aspects. Table 2.3 gives an overview of the papers and categories. Note that, when authors optimize their problem using a deterministic approach, and evaluate their results stochastically, we categorize them as deterministic.

2.5.1 Patient arrivals

Variability in arrivals is a characteristic aspect of problems in an online setting.

A deterministic approach towards patient arrivals is a relevant approach for offline planning decisions, as all information with respect to arrivals for a time period is gathered before a decision is made. We found 33 papers that model patient arrivals deterministically, which are all on capacity planning or offline planning.

A stochastic approach towards patient arrivals is a relevant approach for online planning decisions, as future arrivals are still unknown. In this case, patient arrivals are often represented by a Poisson distribution. A second area in which stochastic patient arrivals are considered are capacity planning problems, as patient arrival information is not yet known at this hierarchical level and has to be forecasted. In all papers found, an average of the historic data is used for

Table 2.3 Variability aspects

Variability aspect	Deterministic approach	Stochastic approach
Patient arrivals	[16, 17, 34, 40, 55, 66–68, 75, 76, 92, 103, 106, 143, 146, 148, 175, 177, 181, 197, 198, 222, 223, 241, 254, 262, 263, 265, 269, 273, 277, 313, 325]	[15, 25, 50, 69, 82, 88, 125, 147, 153, 156, 172, 178, 196, 232, 235, 236, 239, 245, 250, 266, 303–306, 315, 320, 328]
Appointment durations	[15–17, 25, 34, 40, 55, 66–69, 75, 76, 82, 88, 92, 103, 125, 143, 146–148, 153, 177, 181, 197, 198, 218, 232, 236, 239, 241, 245, 254, 262, 263, 265, 266, 269, 273, 277, 303, 305, 306, 313, 325, 328]	[50, 106, 156, 175, 178, 196, 222, 223, 235, 250, 304, 315, 320]
Resource capacity	[15–17, 25, 34, 40, 50, 55, 66–69, 75, 76, 82, 88, 92, 103, 106, 125, 143, 146–148, 153, 156, 172, 175, 177, 178, 181, 196–198, 218, 222, 223, 232, 235, 236, 239, 241, 245, 250, 254, 262, 263, 265, 266, 269, 273, 277, 303–306, 313, 315, 320, 325, 328]	
Care pathway	[15–17, 34, 40, 55, 66–69, 75, 76, 82, 88, 92, 103, 106, 125, 143, 146, 148, 153, 172, 175, 178, 181, 196–198, 218, 222, 223, 232, 235, 236, 239, 241, 245, 250, 254, 262, 263, 265, 266, 269, 273, 277, 303, 305, 306, 313, 315, 320, 325, 328]	[25, 50, 147, 155, 156, 177, 304]

this. In total, we found 27 papers that model patient arrivals stochastically.

In some cases, the robustness of a capacity planning or offline planning approach is evaluated by simulating the capacity-to-patient assignment in a stochastic environment, under the restrictions provided by the (deterministically determined) capacity planning [34, 262] or offline planning [16, 17, 40, 175].

2.5.2 Appointment durations

Variability in appointment durations is a characteristic aspect in problems with time constraints or objectives concerning waiting time, idle time or overtime.

A deterministic approach towards appointment durations is a relevant ap-

proach for planning problems where appointment durations have a low variance, such as check-ups, where no patients and staff are directly involved, such as laboratory and sterilization processes [263], or where variations in appointment durations do not cause significant problems later on the day or in consecutive care stages. We found 46 papers that model appointment durations deterministically.

A stochastic approach towards appointment duration is a relevant approach for problems with time constraints or objectives concerning waiting time, idle time or overtime, and in particular for flow-shop or mixed-shop systems with multiple appointments per patient on one day. In these systems, bullwhip effects can occur due to the interrelatedness of appointments, where delays in one step induce enlarged delays in all further downstream steps. This happens for example if all patients have a fixed order of adjacent scheduled appointments and the first appointment of the first patient takes longer than expected. We found 13 papers that include appointment durations stochastically, in most cases by simulation modeling.

From the papers with flow-shop or mixed-shop systems where patients have multiple appointments on one day and the objectives concerning waiting time, idle time or overtime apply, only two papers model appointment durations using an empirical distribution [153, 178]. The others model them deterministically [67–69, 181, 266, 269].

2.5.3 Resource capacity

Variations in resource capacity can be due to longer term reasons such as sabbaticals or parental leave, or shorter term events such as illness or machine breakdowns. Especially for multi-disciplinary care, variability in the resource capacity of a discipline can have a large effect on the utilization of capacity of interrelated disciplines [267], especially in flow-shop systems where all appointments have to be rescheduled if the first is canceled.

A deterministic approach towards resource capacity is a relevant approach for planning problems where enough capacity is available or where resources can easily be replaced. In these cases, if it turns out that there is more or less capacity available than planned, the appointment can still take place and subsequent appointments are not affected. We found 61 papers that model resource capacity deterministically.

A stochastic approach towards available resource capacity is a relevant approach for all problems where capacity is scarce and where there is time between the moment of decision making and the actual moment to which the decision applies. This holds especially for capacity planning problems, since decisions are made while specific information with respect to resource capacity is not yet known.

We found only two papers in which the total available capacity is modeled stochastically, taking into account longer term capacity variations [34, 146]. [34] model the absence of doctors due to holidays and illness based on historical data, and these capacity variations influence the access times of patients to a large ex-

tent. In [146], the total available resource capacity for a certain period can vary, but is known at the moment that the allocation of the capacity is determined. Capacity fluctuations on shorter term may have a considerable impact on the continuity of care (e.g., when treatments take several weeks or months, during which care professionals can decide to take days off) and on consecutive appointments in a carousel (which all have to be canceled if there is no capacity available for the first appointment). We have found no papers in multi-disciplinary care that take shorter term capacity variations into account.

2.5.4 Care pathways

Variations in care pathways are most prevalent in problems where a diagnosis is involved or in long treatments, where the length or intensity of the treatment depends on the patient's progress.

A deterministic approach towards care pathways is a relevant approach for planning problems with fixed care pathways, as all information on the care pathway is gathered before the decision is made. In literature, care pathways are most often assumed to be known at the moment of arrival: we found 55 papers that model care pathways deterministically.

A stochastic approach towards care pathways is a relevant approach for problems where changes in the care pathway involve a different amount of care or care from different resources on the short term, as this can influence the continuity of care and the capacity utilization. Incorporating variable care pathways can be beneficial for treatment continuity, for example to ensure that capacity is not fully booked when treatments are likely to take longer than expected. We found 7 papers that take into account that care pathways are still uncertain at the moment of arrival, namely the length of stay [25, 155], the required appointments [50], or the routing of patients to the next care providers [147, 177, 304].

2.5.5 Conclusions and further research

Variability in arrivals, care pathways and available capacity are characteristic aspects of problems considering capacity planning and capacity-to-patient assignment in an online setting.

In the literature, variability in care pathways and resource capacity are hardly taken into account, which might cause the resulting planning solutions not to be robust in practice. Especially planning models where care pathways can change during the course of the care pathway would benefit from taking variability into account, in order to ensure continuity of care without delay. As changes in care pathways require quick adaptations of the planning, possible methods include fast approximation algorithms. This offers an interesting direction for further research.

With respect to the resource capacity, short term fluctuations are essential to take into account since they may have a considerable impact on the continuity of the treatment and on consecutive appointments in a flow-shop. Dealing

with short term capacity variability can be done by anticipating for last minute changes in capacity planning or capacity-to-patient assignment, and by rescheduling doctors or therapists when a last minute change occurs. For the latter, fast solution approaches are essential.

Variability in appointment durations has a large impact on problems with time constraints or time-related objectives. These planning problems would therefore benefit from taking variability in appointment durations into account, because otherwise bullwhip effects may occur. However, in literature, appointment durations are often modeled deterministically, which may result in large waiting times and inefficiencies in practice. When variability in appointment durations is taken into account, it is most frequently analyzed using a simulation approach [156, 315, 320]. In contrast, optimization is not often reported upon. A promising direction for further research would be to use stochastic modeling or robust programming such that care systems can be optimized while taking variability in appointment durations into account.

As approaches for taking variability into account in multi-disciplinary systems might become too computationally involved, researchers can opt for approximations or a deterministic optimization approach combined with a stochastic evaluation, such as a sensitivity analysis with multiple scenarios.

2.6 Applicability and generality

For hospital managers and scientists it is important to know which planning approach is relevant to their situation. Hospital managers want to know if a solution that is proposed in the literature, worked in hospital practice, and whether it will also work in their specific hospital. Scientists want to know under what constraints a solution methodology works, whether the approach is also feasible for other departments, hospitals, or industries with similar characteristics, and whether the performance of that approach is better than other known approaches. To increase the practical applicability, a comparison of the proposed approach with the current hospital practice can be made, using historical data from a hospital. To increase generality in a wide range of healthcare institutes, not only the specific parameter settings of one hospital, but also a wide range of other possible parameter settings reflecting a wide range of hospitals, can be taken into account [174]. To increase scientific relevance, a comparison of the performance of the proposed approach with the performance of relevant approaches in the literature can be made. In order to make a good comparison, numerical experiments based on an extensive dataset need to be provided in the paper [298]. This is also relevant to hospital managers, as they want to implement the best performing solution in their practice.

We first analyze the practice perspective of the approaches in Section 2.6.1, whereupon the literature is evaluated on a scientific perspective in Section 2.6.2. Finally, Section 2.6.3 provides conclusions and opportunities for further research.

2.6.1 Practical perspective

Table 2.4 shows the use of datasets by researchers in the multi-disciplinary research field. Note that some papers did not present numerical experiments, and are therefore not included in the table. The majority of researchers use case studies and real-world data to show the applicability of their research to healthcare practice. This is in line with the findings of [32], who observe that real-world data is generally preferred over theoretical data.

Table 2.4 Datasets

Data source	References
Generated data	[16, 25, 76, 146–148, 197, 198, 232, 265, 269, 273, 277, 305, 313, 325, 328]
Historical data	[15–17, 25, 34, 40, 50, 66–69, 75, 82, 88, 92, 106, 143, 153, 155, 156, 172, 175, 177, 178, 181, 196, 218, 222, 223, 235, 239, 241, 245, 250, 262, 263, 266, 269, 303–306, 309, 313, 315, 320, 325]

Most authors consider multi-disciplinary clinic data from a single hospital or a single division within a hospital. This shows a connection of the proposed solution methods with hospital practice. However, this also comes at a risk of optimizing an approach towards a single problem setting, and makes it hard for other healthcare practitioners to evaluate whether the approach will be beneficial to their hospital settings. To show their multi-disciplinary planning approach is suitable for a wide range of healthcare institutes, [67] test their approach to datasets from two hospitals. Next to this work, we found no other papers that include a comparison between multiple hospitals.

Given that the OR literature reports on many promising approaches, with high benefits for practice, the number of papers found in this review that mention implementation in practice is rather low. This is similar to the implementation rate of OR approaches in mono-disciplinary environments [3]. Therefore, for multi-disciplinary environments it is even more important that a planning solution is clear, explicit, and easily manageable.

When applying OR approaches in practice, two implementation directions can be chosen. First, interventions can be adopted by the hospital without major changes to existing information systems, such as a new planning rule, staff schedule, or priority system [175, 178]. These interventions are often the result of an extensive analysis where the effect of multiple interventions was assessed. Second, planning approaches can be implemented in the existing information systems [40, 69]. However, this type of implementation regularly takes more time and often requires considerable investments from the hospital side. A reason for this is that system changes are bounded by the possibilities of the existing information system(s) and that it requires the available data to be available. Although the possible impact of interventions that are implemented into existing information systems might be larger, it is harder and more costly to achieve than

a successful implementation of simple, hands-on interventions.

Table 2.4 shows that 27% of the papers in this review use generated data. One third of them combines the use of generated and real life data, to show both the practical perspective, as well as some theoretical results. The combination of generated data and real life data is used in two ways. First, generated data can be used to show the performance and optimality of a proposed method, whereafter real life data is used to solve a practical problem [16, 25]. Second, generated data can be used to compare the performance of the proposed method with well-known approaches in the literature, whereafter real life data is used to solve a practical problem [269].

2.6.2 Scientific perspective

Multi-disciplinary appointment planning research is relevant in both the OM/OR and the Medical field. Therefore we assess the scientific perspective in both fields.

Operations Management/Operations Research field

For OM/OR researchers, it is relevant to be informed about the technical details of the developed approach, the modeling novelties, and the performance of the approach. Where the first two items are structured in a generic way and well documented in most papers in this review, the documentation on the performance of methods has no generic structure.

First, the performance of a model can be evaluated by showing that an approach provides the optimal solution, or by showing that for smaller instances an optimal solution is derived (e.g., [16, 25]). This way, researchers show that their approach gives (near-)optimal solutions.

Second, the performance of a model can be compared to the current practice in a partnering hospital (e.g., [17, 175]). This way, researchers show that their approach results in an improvement for their partnering hospital.

Third, the performance of a model can be compared to the performance of known solution approaches present in the literature. We found 2 papers that compared their solution methodology to already known solution approaches [269, 313]. A reason for a low comparison rate is the high variation in problem settings of the multi-disciplinary appointment planning problem. In Section 2.6.1 we saw that many authors perform a case study. Therefore, many authors formulate their problem according to a practical situation which they encountered in the hospital they work with, which creates a high diversity in problem settings. To enable the comparison of approaches, generic multi-disciplinary problem settings should be agreed upon by researchers, which can be extended by authors to create the specific setting of their collaborating health institution. A second reason for a low comparison rate, is the lack of benchmarking instances in multi-disciplinary appointment planning. Although some authors mention that their multi-disciplinary planning datasets are available for other users, these are not widely used, which forces researchers to analyze the performance of a solution

approach in multi-disciplinary appointment planning with their own dataset. The lack of multi-disciplinary appointment planning benchmarking instances is in line with the lack of benchmarking instances in healthcare scheduling in general. Only a few benchmark sets are known for healthcare scheduling problems, such as a patient admission scheduling set [60], and a surgery scheduling set [174]. Nurse scheduling is an exception, as many nurse scheduling benchmark sets are available, and most authors benchmark their approaches against the existing literature.

The papers that included a comparison with known approaches, perform this comparison in different ways. [269] use generated instances to compare their multi agent tabu search approach with a well-known genetic algorithm, although both approaches are implemented in different coding languages. [313] first compare their heuristic with an optimal approach to show their approach often derives optimal solutions for small problem instances, and good solutions for real life instance in a reasonable amount of time. Furthermore, they find that their approach results in a lower computation time than the computation time reported upon in another paper covering a similar problem.

Medical field

The majority of medical research papers found on multi-disciplinary care systems focuses on (medical) outcomes, the involvement of operations research techniques in the design of these systems is rarely reported upon. Novel approaches that improve the quality of care, quality of work, and the efficiency of processes are of interest to medical researchers. As multi-disciplinary systems are increasingly introduced, an efficient organization of these systems is of high value.

Multiple multi-disciplinary systems are reported upon in the medical literature. Examples are the implementation of a carousel for Huntington's disease [301, 302], a rapid access clinic for breast cancer [14], and a multi-disciplinary epilepsy clinic, in which a carousel with consultations with multiple as well as single clinicians are planned [107].

Although the majority of the multi-disciplinary system papers in the medical literature only focus on outcomes and not on the organization of care, two exceptions are present [108, 176]. Both of these papers, which comment on the implementation of multi-disciplinary systems and highlight the predicted outcomes and impact, have an OR equivalent besides the medical paper ([262] and [175] respectively). This way, results are disseminated both to the OM/OR researchers, and to the medical researchers.

2.6.3 Conclusions and further research

Planning of multi-disciplinary care is applicable to many healthcare areas. Therefore, most studies derive their problem formulation from a real-world case, and test their approach in a case study of an existing healthcare setting.

To increase the practical perspective, and to convince healthcare managers

that an approach is suitable for a wide range of health care institutes, researchers can apply their approach to datasets from multiple hospitals. Besides comparing multiple hospitals, authors may already be able to show their approach is generic by analyzing multiple departments within one hospital. This is especially useful when these departments have different characteristics, such as differences in patient load, throughput time, and capacity.

To create impact with research results in a healthcare environment, attention should be paid to the dissemination of both the theoretical as well as the practical results [44].

2.7 Conclusions and open challenges

This section presents the conclusions of this review in Section 2.7.1. Following from the conclusions, several open challenges and research opportunities are identified in Section 2.7.2.

2.7.1 Conclusions

This chapter provides a review of the literature on multi-disciplinary planning in healthcare. We evaluated all prescriptive studies on this topic based on several classification fields, including the application area, the decision delineation/hierarchical level, system characteristics, the incorporation of uncertainty, and the theoretical and practical perspective.

Multi-disciplinary systems are more and more observed in the medical context. They are characterized by their low no-show rates, since patients do not intend to risk missing multiple appointments in a row. Most care systems work with pre-scheduled appointments, except for the multi-disciplinary walk-in setting at the emergency department. The involvement of multiple disciplines with often limited availability and time restrictions between appointments make the planning problems complex and highly constrained. Furthermore, often a mix of objectives is considered, since not only the system as a whole should be optimized, but the individual disciplines also require a good performance.

Despite the characteristic aspects of multi-disciplinary planning, it is hard to differentiate multi-disciplinary planning literature from general appointment planning literature without an in-depth literature search, as only few researchers specifically mention the multi-disciplinary nature and its specific characteristics in their work. Therefore, this research aims to give researchers in this field an overview of the available literature in this research area. However, we encourage future researchers to clearly state the characteristics of their work, by including whether they observe a system with multiple interrelated appointments per patient and servers of multiple disciplines or facilities.

Many studies have focused on offline capacity-to-patient assignment, in which all patient demand is known. However, our experience is that we often encounter situations in practice in which patients are scheduled in an online manner, at the

moment of arrival. Online planning is frequently studied in simulation settings, where several scenarios are evaluated. However, further research is needed in the optimization of online planning systems where variability is taken into account.

Most multi-disciplinary appointment planning research has a strong focus on the impact in practice. Many studies include some sort of case study, which may imply collaborating with a healthcare institute and the testing of OM/OR approaches in practice. However, not many manuscripts include details about actual implementation in the practical setting. Furthermore, the generality and scientific relevance of the work of most reviewed papers is questionable, as only one case study is evaluated.

To analyze a multi-disciplinary system, the chosen optimization and evaluation method should be able to deal with a large state and decision space in reasonable time. Therefore, many researchers opt for heuristics and simulation studies. Furthermore, the motivation for organizing the care in a multi-disciplinary fashion should be taken into account, as this is a driver for the performance measures to include in the models. When many variable aspects are involved in a planning decision, researchers may consider designing specific planning solutions that can deal with variability, to avoid bullwhip effects.

2.7.2 Open challenges

Since multi-disciplinary appointment planning is an emerging field in medical practice, and therefore in healthcare optimization research, we expect increasing attention for this research field from researchers in the near future. From this review, multiple future research directions can be derived:

1. This review showed cross-relations in optimization and design questions of different healthcare applications that were solved in isolation. It is an open challenge to develop general approaches for systems with similar characteristics in multiple medical contexts.
2. Planning problems in multi-disciplinary care systems are often complex and need to be robust against all kinds of variability. It is an open challenge to facilitate in the high need for stochastic optimization methods that can deal with a large state and decision space in reasonable time.
3. Section 2.5 showed a lack of robust planning solutions with respect to variability in treatment requirements, resource capacity and appointment durations. It is an open challenge to develop approaches that incorporate multiple of these uncertain variables.
4. Online planning systems, for which many health care applications exist, are most frequently analyzed using evaluation methods. It is an open challenge to optimize online systems, where many future scenarios need to be taken into account, for example by means of stochastic programming, dynamic programming or robust optimization.
5. From a healthcare perspective, it is an open challenge to evaluate and optimize the alignment of staffing schedules to multi-disciplinary clinic appointment schedules, multi-disciplinary team meetings, and combination

appointments in a static and dynamic setting.

6. As healthcare institutes deliver care from a patient perspective, industry solutions need to be adapted from a system or end-user focus to a patient (product) focus. It is an open challenge to analyze flexible flow-shop systems for systems that focus on patient centeredness, and personalized diagnostics and treatment.
7. Only few researchers analyzed open-shop planning, where performance is assessed on the combination of access time and direct waiting time or throughput time. It is an open challenge to combine both time measures in optimization approaches, as two time scales are involved.

To maximize the chances of actual implementation, researchers should closely collaborate with the multi-disciplinary team and be aware of their motivation for organizing care in a multi-disciplinary fashion. This will increase the focus on patient centeredness as well as the development of clear and simple solutions.

In sum, since health care systems are more and more organized in a multi-disciplinary way, and various health care applications share common characteristics and underlying models, cross-relations within different applications can enrich the knowledge on multi-disciplinary care planning solutions. With this review, we encourage researchers to combine the insights and methods from cross-related applications to lift the planning of multi-disciplinary care to a higher level.

PART 2

diagnostics

Why Wait?
Organizing Integrated Processes in Cancer Care

Chapter 3

A.G. Leeftink, R.J. Boucherie, E.W. Hans, M.A.M. Verdaasdonk, I.M.H. Vliegen, and P.J. van Diest. Batch scheduling in the histopathology laboratory. *Flexible Services and Manufacturing Journal*, https://doi.org/10.1007/s10696-016-9257-3, 2017.

Chapter 4

A.G. Leeftink, R.J. Boucherie, E.W. Hans, M.A.M. Verdaasdonk, I.M.H. Vliegen, and P.J. van Diest. Predicting turnaround time reductions of the diagnostic track in the histopathology laboratory using mathematical modelling. *Journal of Clinical Pathology*, 69(9):793-800, 2016.

CHAPTER 3

Optimization of pathology processes - a heuristic approach

3.1 Introduction

Histopathology and anatomic pathology laboratories aim to deliver timely diagnoses to patients. Among others, the laboratories deliver rapid diagnoses during surgeries and fast diagnostics for patients suspecting to have cancer. In this context, high quality care within the shortest possible time is expected from the laboratories. Consequently, employees experience a high work pressure. Therefore, challenges exist regarding both employee workload and turnaround times all over the world [47, 212].

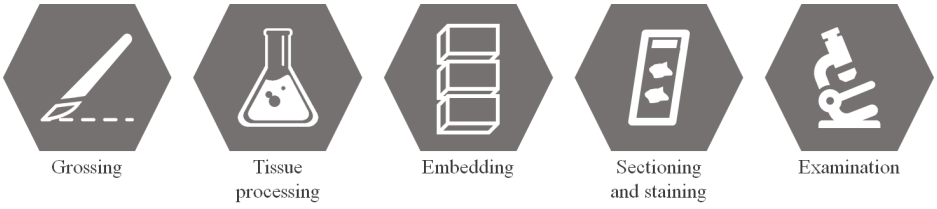
This work is motivated by the histopathology laboratory of UMC Utrecht, where tissue processors (batching machines) are in the middle of a labor intensive multiple stage process chain. This chapter considers the scheduling of tissue samples over the various histopathology activities, from which one activity is executed by multiple batching processors. Chapter 4 presents a case study in the histopathology laboratory of UMC Utrecht. The aim of these studies is to deliver fast reports for the patients and to create a leveled workload for the employees. Or, in other words, we aim to provide increased speed of diagnostics and reduced work pressure.

3.1.1 Histopathology processes

Histopathology processes are complex processes [45]. The processes can be divided into five main activities: grossing, tissue processing, embedding, sectioning and staining, and examination [176], as shown in Figure 3.1. In the grossing stage, tissues are trimmed in representative parts by a technician, and put into cassettes. In the automated tissue processing stage, the tissue in these cassettes is fixated and dehydrated using various chemicals. This process takes up to 12 hours depending on the tissue size, and multiple cassettes are batched during this process. After tissue processing, the tissues are embedded in paraffin wax by a technician, to be sectioned in very thin sections by another technician. When these sections are put on slides, the slides receive a staining using an automated stainer which can be continuously loaded. Hereafter, the residents and patholo-

gists can examine the slides under the microscope or using digital examination. All stages consist of multiple employees (single-unit parallel processors), except for the tissue processing stage. Here, parallel batch processors are to be scheduled with large processing times compared to the other stages. All jobs have equal routing through all stages, and all jobs have a due date reflecting their priority.

Figure 3.1 Histopathology process flow



3.1.2 Performance indicators

In order to provide increased speed of diagnostics and reduced work pressure in the histopathology laboratory, we consider the workload and turnaround time as main performance indicators.

A balanced distribution of workload is important for employee well-being and efficiency in the laboratory. In high workload situations, employees have to put in much effort to complete all their tasks within certain time limits [271]. A high workload has several detrimental effects for both the employee as well as the organization, such as fatigue, psychological distress, increased absenteeism, and reduced quality of work [102]. Furthermore, high workload situations need to be followed by longer periods of employee recovery compared to regular working and recovery situations [202]. Therefore, it is important to limit the time employees have to perform tasks in a high workload environment. Hence, the laboratories should strive for a balanced distribution of workload over the day. This research aims to level the workload perceived by staff. We do so by minimizing the inventory peaks, as the inventory is seen by staff as the amount of work that needs to be done. Also, the exact amount of workload cannot (rapidly) be assessed by staff, as jobs look very similar but may require very different processing times. When striving for a leveled workload, the maximum inventory should be minimized.

The turnaround time (TAT) is a widely used quality measure in pathology, and is often used as the main performance indicator of the histopathology laboratory, together with quality indicators such as diagnostic accuracy [212, 230]. Patients awaiting their diagnosis experience high anxiety and uncertainty levels, especially in cancer care [179, 231]. Thus, patients should be provided with a timely diagnosis, and results should be available to their clinicians as early as technically possible [257]. Therefore, laboratory activities need to be scheduled in such a way that the tardiness of jobs is minimized.

3.1.3 Aim of the research

The TAT and workload optimization of a series of processes requires a system-wide approach. In this research we aim to integrally optimize processes by considering all resources involved, and addressing the tactical level of control in addition to the operational [129]. To address both levels, we decompose the problem at hand in the batching problem and the scheduling problem. More specifically, on a tactical level we determine optimal batch completion times in order to spread the workload. To solve this batching problem, we use an (M)ILP approach. Furthermore, on an operational level the jobs are scheduled such that the tardiness of jobs is minimized using adaptions of existing scheduling approaches. In Section 3.3 we show that both these sub-problems are NP-hard.

Although this chapter is motivated by the histopathology laboratory, scheduling of multi-stage process chains with batch processors is also relevant in other systems in healthcare, and in manufacturing environments. A healthcare example can be found in the sterilization plant, where centrifuges are an essential part of the processes. In manufacturing, an example is a ceramic plant, where pottery has to be baked in an oven.

In sum, our contribution is threefold. First, we optimize a new 3-stage system, where batching is included in the second stage. This NP-hard problem is not only relevant for health care, but also applicable in other industries. By considering this system, this chapter contributes to the scarce literature on hybrid flow-shops with parallel batching. Second, we develop a new, novel solution method which addresses both the tactical level as well as the operational level. Third, the practical applicability of this method is demonstrated with a real-life case study of a large academic hospital in the Netherlands in Chapter 4.

This chapter is organized as follows. Section 3.2 describes the literature, followed by the formal problem description in Section 3.3, and the batching and scheduling models in Sections 3.4 and 3.5. Section 3.6 presents the experiment results. Section 3.7 gives conclusions and opportunities for further research.

3.2 Literature

In the histopathology laboratory, parallel batch processors are part of the process chain. Chapter 2 showed laboratory processes can be represented by a flow-shop system. In this chapter, we specifically consider a 3-stage hybrid flow-shop with multiple identical parallel batching machines in the second stage. In a hybrid flow-shop (HFS), at least one stage in the flow-shop has multiple machines, in a HFS with parallel batching (HFPB) some of the machines considered can process multiple jobs simultaneously [11]. This system is known from the process industry [122, 131, 203]. In Section 3.2.1 we first describe the HFS scheduling literature, since this is the core problem under consideration, followed by the literature on the HFPB in Section 3.2.2, where parallel batching constraints are added to the HFS. Note that in the remainder of this section, unless otherwise stated, we

specifically focus on HFS and HFPB problems that consider scheduling while minimizing completion time, overtime, and tardiness objectives.

3.2.1 HFS

The HFS is well studied in the literature. The HFS is a flow-shop in which in at least one stage multiple machines are available. We refer to Ruiz and Vazquez-Rodriguez [264] and Ribas et al. [256] for extensive reviews of the HFS literature. The HFS literature can be divided in three categories, based on the number of stages: two-stage, three-stage, and k-stage [6]. Only some specific configurations of the two-stage HFS (i.e., $F2||C_{max}$, see Johnson [150]) are solvable within polynomial time. The problem of scheduling the two-stage HFS with parallel machines is NP-complete for most regular scheduling objectives, even in its simplest form with one machine in one stage, and two in the other stage [120, 142].

Frequently used solution methods for HFS scheduling problems are exact methods, heuristics, and metaheuristics [264]. Exact methods are often based on branch and bound techniques and can only be applied to problem instances of small sizes for complexity reasons [183]. Therefore, many authors use heuristics, metaheuristics, or a combination of these techniques to solve HFS and HFPB scheduling problems. Dispatching rules, such as Earliest Due Date (EDD), Shortest Processing Time (SPT), and Longest Processing Time (LPT) have shown to result in good performing solutions for specific optimization criteria in HFS scheduling and batch scheduling and are easily applicable in complex environments [183, 203, 264]. Metaheuristics, such as simulated annealing and genetic algorithms, are able to improve upon these solutions [264].

Besides variations in the number of stages and solution methods, HFS configurations are distinguished based on different objective functions (i.e., flow time or due date based [11]) and constraints. These constraints are added to the general HFS to come closer to real-life problems. Allaoui and Artiba [7] studied the HFS with *availability constraints*, to ensure preventive maintenance activities to be executed. Mirsanei et al. [206] proposed a simulated annealing approach for the HFS problem with *sequence-dependent setup times*. *Multiprocessor task scheduling* in a hybrid flow-shop environment was studied by Engin et al. [96]. They developed a genetic algorithm to solve the HFS with multiprocessor tasks. The HFS with *recirculation* was studied by Bertel and Billaut [33]. They proposed an ILP formulation, and developed a genetic algorithm for solving industrial size instances. *Precedence constraints* in a 2-stage HFS with parallel machines in the second stage was considered by Carpov et al. [54]. A randomized list scheduling heuristic was proposed, together with the examination of global lower bounds. Gupta et al. [121] studied a similar system, but with parallel machines in the first stage, and a single machine in the second stage. They proved both problems are equivalent and NP-hard, and proposed a branch and bound algorithm using several lower bounds and heuristic methods. A combination of these two systems, together with the aforementioned batching requirements, reflects the 3-stage system under review in this study.

3.2.2 HFPB

The extension to the HFS of interest in this chapter is the HFS with *parallel batching*. There are two types of batching known; serial batching and parallel batching. In serial batching, jobs in a batch are processed sequentially. For a recent example of HFS with serial batching, we refer to Ghafari and Sahraeian [111], who proposed a genetic algorithm. In parallel batching, all jobs in a batch are processed in parallel. Only a limited number of studies considered a hybrid flow-shop with parallel batching in one or multiple stages, where batch compositions can differ throughout the stages. Bellanger and Oulamara [28] were the first to study the two-stage HFPB with parallel batching in the second stage with task compatibilities, which they motivated by the tire manufacturing industry. They proposed three heuristics together with their worst-case analyses. Luo et al. [187] considered a two stage HFPB with parallel batching in the first stage, motivated by the processes of a metalworking company. Since the problem is NP-hard, they determined the batches upfront using a clustering algorithm, to reduce the problem complexity. More recently, Rossi et al. [263] studied a two-stage HFPB reflecting a hospital sterilization department. They recommend closing batches before completion, for example as a function of the elapsed time or by fixing the capacity threshold. The work of Amin-Naseri and Beheshti-Nia [11] is the closest to our problem, since they studied the 3-stage HFPB where batching was allowed in any stage aiming to minimize the maximum completion time. They proposed three two-phased heuristics based on a combination of Johnson's rule, scheduling algorithms for parallel machines, and theory of constraints. Furthermore, they developed a three dimensional genetic algorithm which outperformed their heuristics. However, besides the different objective function, they allowed for jobs to start processing on batching machines after the first job of that same batch already started processing, since only completion times were aligned, which is not applicable to our case.

3.2.3 Conclusions

Concluding, it is known that the HFS problem in which the tardiness is minimized in itself is a complex problem, since it is NP-hard. Furthermore, there is only scarce literature available on the HFS extension with parallel batching, since the complexity of the HFS increases even more by adding these constraints. In addition to the HFPB with a tardiness objective, we encounter a workload leveling objective. To the best of our knowledge, this system, the 3-stage HFPB with parallel batching in the second stage where the intermediate storage has to be kept to a minimum, has not been considered before in the literature and has never been applied in a hospital or manufacturing setting.

3.3 Formal problem description, complexity, and decomposition

We consider a set of G stages, with each stage g having M_g identical parallel resources. All jobs need processing in all stages and can be processed by all resources. Each job j is of a certain job family f, with a corresponding release time r_j. The processing times $p_{j,g}$ are known for each job in each stage, and deterministic, based on the job family. Preemption of jobs is not allowed, and jobs cannot be split over multiple machines in a stage. Transportation times are not included in the model. If they would be included, the only effect is seen in the timing of jobs in the next stages, which would increase with the transportation times. Operator workload is not affected, since transportation is an automated process. There is unlimited intermediate storage available between stages, which should be kept to a minimum. In order to schedule all jobs over all resources, regular working hours are considered. Thus, for all resources, the start time s and end time e for each day are known. Resource breakdowns are not included. In the second stage, parallel batch machines are available. The capacity of each batch b is unlimited. Furthermore, the processing time of a batch p^b equals the largest processing time of all jobs that are assigned to that batch (3.1).

$$p^b = \max_{j \in J^b} p_j \quad \forall b \in B, \tag{3.1}$$

where J^b is the subset of jobs that are assigned to batch b.

The jobs in this system should be scheduled in such a way that the maximum inventory between stages and the tardiness of jobs (i.e., the amount of time jobs are finished after their due date) are minimized. Following the notation of Graham et al. [114], the problem can be described as:

$$FH3B(m_1, m_2, m_3)|p - batch(2), r_j|I_{max}, \sum T_j. \tag{3.2}$$

Here, a three-stage HFPB with m_1 resources in the first stage, m_2 resources in the second stage, and m_3 resources in the third stage is defined. $p - batch(2)$ shows that stage 2 consists of parallel batching machines, and r_j shows that varying release times are considered. The two performance indicators considered are the maximum peak inventory levels (I_{max}) and the sum of the tardiness of all jobs ($\sum T_j$), which are both minimized. An optimal solution is an assignment of each job in each of the stages to a machine or resource, given certain availability, precedence, and time constraints, that minimizes the aforementioned objectives. The mathematical problem formulation can be found in Appendix 3.8.

The problem is unary NP-hard, since the two stage HFS as well as the flow-shop with batching are known to be NP-hard [120, 243]. Two options remain for solving the problem: Complete enumeration over all possible solutions, or heuristics [114]. Due to the large solution space, complete enumeration will be prohibitively time consuming, thus we will focus on heuristics for solving this problem.

A few approaches exist that combine batch size, batch assignment, and batch sequencing decisions [244]. However, these approaches only allow for very small instances, with limited number of resources and jobs [130]. Therefore, in accordance with the literature, we propose to decompose the batching and scheduling decisions in the remainder of this research. Since in practice the batch timing is often determined at a tactical level a few times a year, and the job scheduling is an operational level task, we developed a new decomposition approach in which the batch timing is determined first ('*Batching problem*', see Section 3.4), and the scheduling of jobs second ('*Scheduling problem*', see Section 3.5). Herein, the batching problem aims to minimize the inventory peaks, whilst the scheduling problem aims to minimize the total tardiness.

3.4 Phase 1: Batching problem

The batching problem focuses on scheduling batches while aiming to minimize the inventory peak between the batching and its subsequent stage. This section gives the formal problem description, including the definitions, goal, and approach.

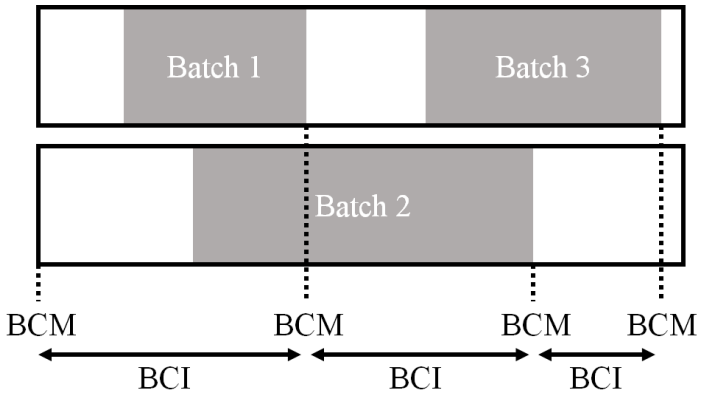
3.4.1 Problem description

Consider a three-stage HFPB with multiple parallel batching machines in the second stage. When a parallel batch processor is followed by labor-intensive processes, the highest inventory peaks, and therefore peaks in workload, occur at the moment a batch processor is finished and all jobs become available for the next stage. Since the batch processing time depends on the size of the jobs in the batch, the batching moments in relation to their output should be controlled in order to equally spread the inventory. Therefore, the batching problem determines the timing of the batches, while the minimum interval between two subsequent batches is maximized.

The work of Van Essen et al. [97] is the closest to our approach. They developed several solution methods to minimize the interval between completion times of scheduled surgeries by optimizing their sequence. Their work reduces the expected waiting time of emergency surgeries, which may start at the aforementioned completion times. They proved this problem is strongly NP-hard for two or more operating rooms [97]. However, their aim is to minimize the maximum interval, whilst we want to maximize the minimum interval. Furthermore, they assumed the surgeries were already assigned to fixed operating rooms. We consider the more advanced case, where the batches have to be scheduled over multiple machines. However, this comes against a cost of an increased solution space and additional decision making, which makes the problem even harder to solve.

The moment that a batch is finished is referred to as batch completion moment (BCM). The interval between two subsequent BCMs is defined as the batch completion interval (BCI), see Figure 3.2. The length of the BCIs depends on

Figure 3.2 Batch Completion Moments (BCMs) and Batch Completion Intervals (BCIs) for a 2-machine problem



the assignment, sequence, and timing of the batches. Since the batching problem is considered at a tactical level, we assume no information on future job arrivals is available.

3.4.2 Definitions and goal

Our aim is to find a cyclic batching schedule, at a daily level. For a day, we consider the set of B batches each of a given batch type t, where each batch of a certain type has a corresponding batch type processing time p_t. All jobs with a processing time lower than the batch type processing time (i.e., $p_{j,2} \leq p_t$) are eligible for a batch of this batch type.

To spread the output of batches during the day, we aim to equally spread the BCMs over the day, such that peaks in workload in the subsequent stage are minimized. Taking into account the expected load of one batch, we want to maximize the interval between two subsequent batches, and between subsequent batches with the same load. We hypothesize that the smallest interval accounts for the highest peak in workload, and is therefore our main objective: the maximization of the smallest BCI.

If multiple batches of the same batch type are scheduled, the minimum batch completion interval is maximized per batch type as well. This ensures that jobs of a certain job family are processed more equally spread over the day, which is especially important when weights are added to having certain job families in inventory, as in the histopathology laboratory.

3.4.3 Approach

To determine the optimal batching schedule, we formulated a Mixed Integer Linear Program (MILP). Despite the NP-hardness of the optimization problem, we can solve real-life instances using this mathematical program. In the MILP, batch sequencing, batch timing, and batch-machine assignment constraints are

included. This way, the BCIs can be determined, using the sequence in which all batches are finished by taking the interval in between subsequent batches. We first introduce some additional notation. Hereafter, the objective and constraints are given.

Notation: $OBJ1$ and $OBJ2_t$ are the two objective variables, where the first one is the minimum overall batch completion interval, and the second the minimum batch completion interval per batch type t. Let $X_{b,m}$ be 1 iff batch b is assigned to machine m, and 0 otherwise. Let $Y_{b,b'}$ be 1 iff batch b ends before batch b'. The sequence in which all batches finish is stored by the position of each batch. Let P_b be the position of batch b in this sequence. S_b and C_b refer to the starting time and end time of batch b respectively. Finally, let $\mathcal{M}$ be a sufficiently large number.

Objective:

$$\max \alpha OBJ1 + \beta \sum_{t \in T} OBJ2_t \tag{3.3}$$

Constraints:

$$\sum_{m} X_{b,m} = 1 \qquad \forall b \in B \tag{3.4}$$

$$\sum_{b'} Y_{b',b} + 1 = P_b \qquad \forall b \in B \tag{3.5}$$

$$Y_{b,b'} + Y_{b',b} = 1 \qquad \forall b, b' \in B, b < b' \tag{3.6}$$

$$P_{b'} - 1 + \mathcal{M}^1 \cdot Y_{b',b} \geq P_b \qquad \forall b, b' \in B \tag{3.7}$$

$$S_b + p^b = C_b \qquad \forall b \in B \tag{3.8}$$

$$S_b \geq s \qquad \forall b \in B \tag{3.9}$$

$$e \geq C_b \qquad \forall b \in B \tag{3.10}$$

$$S_{b'} + \mathcal{M}^2 \cdot (2 - X_{b,m} - X_{b',m}) \geq C_b - \mathcal{M}^2(1 - Y_{b,b'}) \qquad \forall b, b' \in B, m \in M \tag{3.11}$$

$$C_{b'} - C_b + \mathcal{M}^3(1 - Y_{b,b'}) \geq OBJ1 \qquad \forall b, b' \in B \tag{3.12}$$

$$C_{b'} - C_b + \mathcal{M}^3(1 - Y_{b,b'}) \geq OBJ2_t \qquad \forall b, b' \in B_t, t \in T \tag{3.13}$$

$$\text{all variables} \geq 0 \tag{3.14}$$

The objective of the ILP is a weighted sum of the two objectives mentioned, i.e., maximize the minimum batch completion interval and maximize the minimum interval between the completions of two batches of the same type (3.3). Each batch should be assigned to exactly one machine (3.4). The position of

a batch in the sequence equals the number of batches finished before this batch added with one (3.5). For example, if a batch is the third one to finish, there were already two batches that finished before him, thus the position in the sequence $P_b = 2 + 1 = 3$. A batch b is either scheduled before a specific other batch b', or after that same batch b' (3.6). Cycles in the positioning are not allowed, thus the position of batch b should be strictly less than the position of batch b', if batch b is scheduled before batch b' (3.7). The completion time equals the starting time of a batch plus its processing time (3.8). A batch starts processing after the machine starting time (3.9), and finishes before the machine end time (3.10). The completion time of a batch should be smaller than the starting time of a successive batch scheduled on the same machine (3.11). We want to find the minimum batch completion interval between all batches (3.12) and between the batches from each batch type (3.13).

Let B be the number of batches, T the number of batch types, and M the number of machines. Then, the MILP consists of $5B+(M+T+2,5)B^2$ constraints and $1 + T + (3 + B + M)B$ variables, from which $(1 + B + M)B$ integer and $1 + T + 2B$ continuous. Thus, for real life instances, with a maximum of 4 machines, 12 batches, and 3 batch types, the batching problem consists of 1428 constraints and 232 variables.

3.5 Phase 2: Scheduling problem

The scheduling problem focuses on jointly scheduling all jobs in all stages, while aiming to minimize the tardiness of jobs. This section gives the formal problem description, including definitions, the goal, and approach.

3.5.1 Problem definition

The optimization problem in which jobs are scheduled on a single-machine in order to minimize the total tardiness, is proven NP-hard [93]. The multi-machine case with multiple stages increases the computational complexity of this single-machine problem, and therefore is NP-hard as well. Therefore, exact scheduling approaches, for example based on Gupta and Karimi [122], can only be developed to solve small instances of the resource assignment and scheduling problem. This exact approach is formulated as an ILP, by fixing the batch times in the ILP formulation from Appendix 3.8. However, due to the complexity of the problem, solving this adapted ILP still takes more than a week for real life instances, while in the histopathology practice a solution should be generated in less than 10 minutes.

As an approximation alternative, we consider a list scheduling algorithm. A list scheduling algorithm is a well known method to multi-machine job shop scheduling [157]. It generates fast solutions, and can easily be implemented in the histopathology practice. Therefore, we propose a list scheduling heuristic to solve the scheduling problem.

3.5.2 Definitions and goal

Our aim is to find a job-machine assignment for a given problem instance. Herein, the batch timing is known, but jobs still need to be assigned to batches. We propose a list scheduling algorithm, in which multiple sequencing rules can be taken into account [292]. We consider the following sequencing rules:

- EDD rule: Arrange jobs based on their due date d_i, and select the earliest due job first.
- SPT rule: Arrange jobs based on their processing time $p_{i,j}$, and select the job with the shortest processing time first.
- LPT rule: Arrange jobs based on their processing time $p_{i,j}$, and select the job with the longest processing time first.

Furthermore, we consider some modifications to these rules, since the due dates and processing times of jobs may be equal for similar jobs:

- EDD-SPT rule: Arrange jobs first based on their due date d_i, if due dates are equal, arrange jobs on their processing time $p_{i,j}$. Select the earliest due jobs first, and, if due dates are equal, the jobs with the shortest processing time first.
- SPT-EDD rule: Arrange jobs first based on their processing time $p_{i,j}$, if due dates are equal, arrange jobs on their due date d_i. Select the jobs with the shortest processing time first, and, if processing times are equal, the earliest due jobs first.

The objective of the scheduling problem is to minimize the tardiness of jobs. However, the maximum inventory level is evaluated as well, being the output of the decomposition approach.

3.5.3 Approach

We designed a multi-phase list scheduling algorithm. In each phase s, it selects a machine m, chooses an unscheduled job j, assigns this job to the earliest available time at this machine, and updates the machine availability and job status. The choice of the machine depends on the availability of the machine. The choice of the job depends on the chosen sequencing rule and the job availability.

In the first phase, the algorithm assigns jobs to batches of the second stage of the HFPB. In the second phase, the jobs are scheduled in the first stages, and the third phase schedules the jobs in the third stage. It might be needed to reschedule the batch assignment in the second phase, since after assignment in the second phase, the jobs might not be finished processing in the first stage before the original batch timing. However, by first scheduling the batches, and thereafter the first stage, jobs that have an earlier due date, but later batch timing, might be assigned to a stage one machine after a job with a later due date but earlier batch timing. This way, the later due job might not unnecessarily be delayed by one batch, and the earlier due job is still processed on time for its own batch.

Table 3.1 Variations in experiment variables

Variable	Interval
M_1	1,2
M_2, B	(1,2), (1,3), (2,2), (2,3), (2,5), (4,3), (4,5), (4,8)
M_3	3,5,7
F	1,2,3
J	10,80,130

3.6 Experiment design

This section describes the experiments that are conducted to analyze the performance of the proposed methods. The impact of the problem size and sequencing rules are evaluated in terms of tardiness and maximum inventory level. We test the batching and scheduling approach on 342 scenarios, as described in Section 3.6.1, and evaluate the performance in Section 3.6.2. Furthermore, a case study is presented in Chapter 4.

3.6.1 Experiment setup

Each experiment spans a one day period of eight working hours. The batching and scheduling problem repeats itself every day. The number of batching machines is set at 1, 2, and 4, with 2, 3, 5, or 8 batches depending on the number of machines. The number of job families is set at 1, 2, and 3, with uniform distributed target throughput times in minutes on the intervals $[320, 500]$, $[540, 950]$, and $[1080, 1800]$ respectively, which corresponds with the histopathology practice. The corresponding discrete batch processing times of each job family are set at 120, 190, and 230 minutes, which reflect the different batch configurations of the histopathology laboratory. The number of non-batching machines is varied between 1 and 2 identical machines in the pre-batching stage, and between 3, 5, and 7 identical machines in the post-batching stage. The number of jobs is set at 10, 80, and 130. The job processing times in minutes in non-batching stages were derived from a uniform distribution on the interval $[5, 15]$, and $[1, 5]$ respectively. All jobs are available at the start of the planning horizon. A summary of all input variables and parameter is given in Table 3.1 and Table 3.2. Note that not all combinations of parameter values are valid, which leaves us with 342 scenarios.

For the batching problem, we derive one optimal solution per problem instance. For each of these instances, 50 scheduling instances are generated using the distributions of Table 3.2. With these instances the effect of the five sequencing rules are analyzed. Preliminary research indicated that 50 replications are needed to obtain relevant results. All experiments are based on random numbers.

All experiments are solved on a HP laptop personal computer with 2GB RAM, using CPLEX 12.6 in AIMMS 4.0 [36, 77].

Table 3.2 Uniform distributed intervals of experiment parameters

Parameter	f=1	f=2	f=3
p_1	[1,6]	[1,6]	[1,6]
p_2	120	190	230
p_3	[1,36]	[1,36]	[1,36]
r_j	[480,480]	[480,480]	[480,480]
$d_j - r_j$	[320, 500]	[540,950]	[1080, 1800]

3.6.2 Performance

We first illustrate the results of the algorithms by discussing the details of one specific experiment instance. Thereafter, overall results on all experiments are presented. The tardiness performance is given in minutes, whereas the inventory performance is given in number of jobs.

Sequencing rules

All scenarios were tested against the five sequencing rules. Table 3.3 shows the results in terms of tardiness and inventory peak for all sequencing rules. The SPT and SPT-EDD sequencing rules both significantly outperform all other rules with respect to the tardiness criterion ($p < 0.01$). For the inventory level, the LPT sequencing rule is significantly outperformed by all other rules ($p < 0.01$), and the SPT rule is significantly worse than the SPT-EDD sequencing rule ($p < 0.01$). The EDD rule performs best on the inventory criterion. Therefore, depending on the performance indicator of interest, a different sequencing rule results in the best performance. However, since both the performance indicators are of our interest, we decided to continue the experiments with the SPT-EDD sequencing rule. The remainder of the experiment results in this section are based on this sequencing rule.

Performance of one experiment

Consider a specific experiment, which represents the situation with one grossing employee, five sectioning employees, four batching machines, five batches during the day, one batch during the night, and 80 jobs divided over three job families. The scheduling model uses the SPT-EDD sequencing rule, since it is the best performing sequencing rule.

In Figure 3.3, the inventory level of the plant during working hours is shown for an instance of this specific experiment. The effects of the SPT-EDD sequencing rule can be observed, since the workload decline is steeper when peaks are higher. Furthermore, one can see that the maximum inventory level equals 22 jobs.

This results in a tardiness of 2133 minutes, which are incurred by 8 jobs, from which 6 jobs are processed during the night. This gives an average tardiness of about 267 minutes ($\approx$ 4.5 hours) per tardy job.

Figure 3.3 Inventory levels for replication 48 of the selected experiment (one grossing employee, five sectioning employees, four batching machines, five batches during the day, one batch during the night, 80 jobs divided over three job families, and SPT-EDD sequencing)

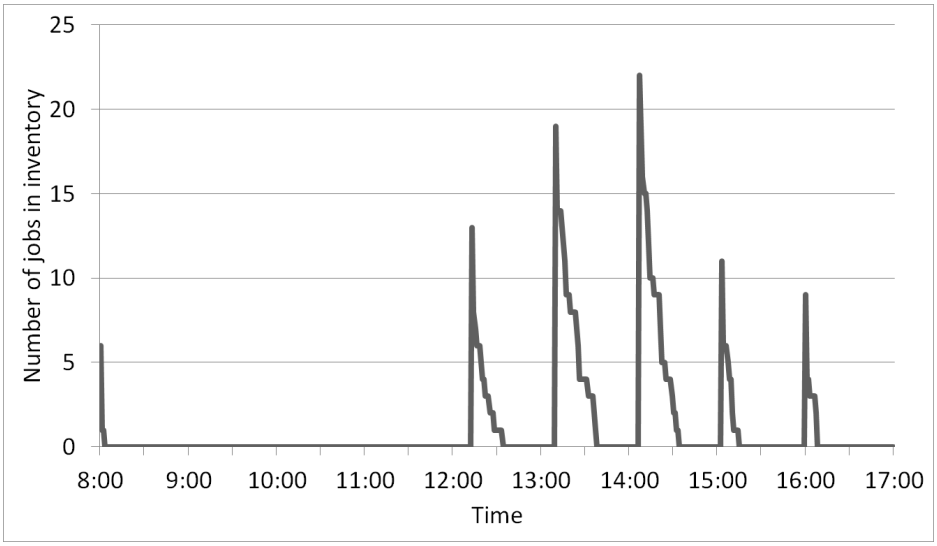


Table 3.3 Sequencing rule experiment results

Seq. rule	$E[\sum T_j]^a$	$\sigma[\sum T_j]^a$	$E[I_{max}]^b$	$\sigma[I_{max}]^b$
EDD	28164	25015	36,7	25,6
LPT	52734	38798	44,3	34,4
SPT	17537	17357	37,5	26,3
EDD-SPT	21638	21825	37,6	26,5
SPT-EDD	17094	17043	37,2	26,0

a in minutes
b in number of jobs

Overall performance

The results of all 342 experiments are shown in Table 3.4, Table 3.5, Table 3.6, and Table 3.7. in these tables, data from multiple experiments are combined to obtain the displayed aggregated results.

Effect of number of non-batching machines An increase in non-batching machines (M_1 and M_3) corresponds with a decrease in tardiness, but not necessarily in improved inventory performance, as shown in Table 3.4. When adding an extra grosser, the peak inventory increases. Since the output of the grossing stage increases, more tissue is processed in the first batches, which causes this inventory peak increase. However, for the third stage, adding extra machines positively impacts the peak inventory, since more jobs can be processed simultaneously, which reduces the inventory at a faster pace.

Table 3.4 Aggregated performance in tardiness in minutes and peak inventory in number of jobs for SPT-EDD sequencing rule

Parameter	Value	$E[\sum T_j]$	$E[I_{max}]$
M_1	1	22985	32,3
	2	10871	42,1
M_3	3	24616	37,9
	5	14409	36,9
	7	11918	36,8
J	10	791	7,4
	80	5539	45,5
	130	28039	61,5
Overall		**17094**	**37,2**

Table 3.5 Average tardiness in minutes per number of jobs scheduled for SPT-EDD sequencing rule

Number of machines	1		2			4		
Number of batches	2	3	2	3	5	3	5	8
10	791	0	0	0	0	0	0	0
80	60959	40575	61007	39071	38976	38899	35416	34359
130	19719	6828	19708	9052	5305	8937	7079	5730

Table 3.6 Average inventory peak per number of family types in minutes for all machine-batch combinations for SPT-EDD sequencing rule

Number of machines	1		2			4		
Number of batches	2	3	2	3	5	3	5	8
1	47	44	47	44	44	44	44	44
2	45	n/a[a]	45	40	n/a	39	28	33
3	n/a	n/a	n/a	34	n/a	34	25	30

[a] For this combination the scheduling problem is infeasible, and therefore no results are displayed

Table 3.7 Average tardiness per number of family types in minutes for all machine-batch combinations for SPT-EDD sequencing rule

Number of machines	1		2			4		
Number of batches	2	3	2	3	5	3	5	8
1	38946	24301	38946	24584	23135	24375	23149	23124
2	41678	n/a[a]	41784	26968	n/a	28432	25461	21154
3	n/a	n/a	n/a	22974	n/a	22740	19824	19619

[a] For this combination the scheduling problem is infeasible, and therefore no results are displayed

Effect of number of jobs Table 3.4 shows the effects of increasing the number of jobs. When more jobs are to be processed, higher inventory levels are present and a higher utilization of resources is derived. Therefore, the tardiness increases, as shown in Table 3.4. Furthermore, a relation between the number of batches and number of jobs can be observed, as shown in Table 3.5. When jobs are processed more spread over the day in multiple batches, the tardiness decreases.

Effect of number of job families Recall that the job families determine the distribution of the due dates and processing times of jobs. Thus, a job of a certain job family has a batch processing time corresponding to that family. Therefore, the complexity of the scheduling problem is expected to increase when including more job families. Where Table 3.6 shows that more job families relate to lower inventory, Table 3.7 does not show a clear relation between tardiness performance and the number of job families. A possible explanation can be found in the characteristics of the job families, which may increase or decrease the possibility to derive a good solution. For example, if a certain job family with a more strict due date is added to an instance, the tardiness will increase compared to the situation where this job family is excluded from the instance.

Effect of number of batching machines and batches Increasing the number of batches run on the different machines has the expected effect of decreasing the peak inventory level, as shown in Table 3.6 and tardiness, as shown in Table 3.7. However, the number of machines, which impacts the timing of the batches, does not show a significant relation with the peak inventory level and tardiness. As shown in Table 3.7, some combinations perform better than others. This indicates that the timing of batches is important to derive a solution with low tardiness and inventory.

Furthermore, Table 3.6 and Table 3.7 show that a tradeoff has to be made between the tardiness criterion and the peak inventory criterion. For example comparing including 5 or 8 batches on 4 machines, one can see that different configurations lead to either a solution with better inventory performance, or a solution with better tardiness.

Summarizing, multiple effects are observed based on the experiments. First, the SPT-EDD sequencing rule performs best according to the tardiness performance indicator. Second, the number of batches, the number of jobs, and therefore the load of the system, has a large effect on the maximum inventory level and on the tardiness of jobs. Third, the number of machines and the number of job families do not show a clear relation with the inventory level and tardiness.

3.7 Conclusions and discussion

The 3-stage HFPB with batching processors in the second stage is a new problem in the literature. We have introduced a decomposition solution method to optimize and prospectively assess the planning and scheduling of batches and jobs, and applied this approach to the processes in the histopathology laboratory. The Phase-1 model includes a novel workload spreading approach, which was based

on a surgery sequencing approach as recently introduced in the literature [97]. Despite its NP-hardness, we could solve real-life instances of this optimization problem to optimality. The Phase-2 model includes a list scheduling algorithm, to ensure practical applicability without compromising the performance.

The results show that in the scheduling model the SPT-EDD sequencing rule performs best in terms of tardiness and peak inventory. Furthermore, it was shown that increasing the number of batches has the expected effect of decreasing the peak inventory level and the tardiness.

In the model, a few assumptions were made. One important assumption is that batches consist of the same amount of jobs. In practice, if two batches of the same batch type are scheduled within a small time frame, only a few new arrivals have occurred, and thus the workload resulting from the second batch will be small compared to the workload resulting from the first batch. However, extensive testing with real life data shows approximately equally sized peaks, which corresponds with the assumption that all batches will induce the same load when supply and demand in the surrounding stages are wisely set. When varying release times are taken into account, this assumption might get violated, since not enough jobs are available for the first batches. In relation to this, the effects of arrival patterns of jobs is of interest. Therefore, further research could be done to include a weighted batch completion interval according to the job arrivals. This might result in even more leveled batch loads, and therefore in a more leveled inventory distribution.

We assumed deterministic processing times for both manual and non-manual tasks. The machines are pre-programmed, and therefore deterministic processing times reflect reality. On the opposite, manual labor work always includes variation, and therefore stochastic processing times seem a more realistic representation of reality. However, there is a large difference in service time of the batch processor (multiple hours) compared to the technicians (a few minutes). Therefore, we expect the effects of variation to be negligible. Furthermore, since decisions are to be made at a tactical level, we expect including stochastic service times of technicians does not have a large influence on the outcomes.

In relation to this, we assumed the use of identical machines with stable processing rates. In practice, employees work at different paces, which would favor non-identical machines in some stages to better reflect reality. Furthermore, employees tend to work harder in the end of the afternoon to finish the last pile of work, when needed, and are mostly willing to work a few minutes in overtime, if that guarantees finishing the last jobs. We did not include these soft deadlines in our model, which might have led to an overestimate of the tardiness and inventory peaks.

This research assumes the workload to be reflected by the maximum inventory level. This relation is especially important for systems with manual activities, such as the histopathology laboratory. However, this objective can be important in manufacturing environments as well, since work-in-process levels should be kept to a minimum [292]. Further research should be executed to evaluate the

actual impact on job satisfaction and the quality of the executed work, since they are hard to prospectively assess using mathematical modeling.

The perceived workload used throughout this research includes the number of jobs, but not the expected processing time, as usual in industry environments. The expected processing times can easily be incorporated in the methods proposed, by introducing a weighted inventory. However, we choose to use the perceived workload in number of jobs, since we concluded, together with the technicians and other laboratory staff, that the perceived workload is mainly influenced by the number of jobs that still has to be done, and not by the minutes of work spent on those jobs. Since jobs look very similar, a technician cannot see upfront whether a job has a high or low processing time, thus the number of jobs is the only practically relevant indication of the perceived workload.

Concluding, this research proposed a decomposition planning and scheduling method in order to reduce the turnaround time and the maximum inventory levels in a three stage HFPB. In the next chapter, we will show practical applicability of this method to a concrete case in a histopathology laboratory of a large university medical center in the Netherlands.

3.8 Appendix I

This appendix presents the model developed for scheduling the 3-stage HFPB. For notation, refer to Table 3.8. The objective and constraints are as follows:

Table 3.8 Notation

Index	Definition
i:	index of job, $i = 1, \cdots, I$
j:	index of machines, $j = 1, \cdots, J$
g:	index of stages, $g = 1, \cdots, G$
b:	index of batches, $b = 1, \cdots, B$
t:	index of time, $t = 1, \cdots, H$
$J_{i,g}$:	set of machines that can process job i in stage g
$J_{i,g}$:	set of machines that are available for processing in stage g
J^{batch}:	set of batching machines
I_j:	set of jobs that can be processed by machine j
$NC_{i,g}$:	set of jobs that cannot be processed by any of the machines that process job i in stage g
$NG_{i,g}$:	next processing stage of order i currently being processed in stage g
G^{batch}:	set of batching stages
G^{final}:	set of final stages

Parameter	Definition
r_i:	release time of job i
r^j:	release time of machine j
$w_{i,g}$:	weight factor of job i in stage g
$p_{i,j}$:	processing time of job i on machine j
pb_j:	processing time of batch b on machine j
d_i:	due date of job i
H:	planning horizon
$\mathcal{M}$:	a sufficiently large number

Variable	Definition
$Z_{i,j} =$	$\begin{cases} 1, & \text{if order } i \text{ is processed by unit } j \\ 0, & \text{otherwise} \end{cases}$
$ZF_{i,j} =$	$\begin{cases} 1, & \text{if order } i \text{ is the first order to be processed by unit } j \\ 0, & \text{otherwise} \end{cases}$
$X_{i,i',g} =$	$\begin{cases} 1, & \text{if order } i \text{ is processed before order } i' \text{ in stage } g \\ 0, & \text{otherwise} \end{cases}$
$Q_{i,j,b} =$	$\begin{cases} 1, & \text{if order } i \text{ is processed in batch } b \text{ on unit } j \\ 0, & \text{otherwise} \end{cases}$
$W_{i,t} =$	$\begin{cases} 1, & \text{if order } i \text{ is in inventory at time } t \\ 0, & \text{otherwise} \end{cases}$
$T_{i,g} =$	Time at which order i starts processing in stage g
$S_b =$	starting time of batch b
$C_b =$	end time of batch b
$D_i =$	delay of job i
$V_t =$	inventory in number of jobs at time t
$V^{\max} =$	maximum inventory after batching stage

Objective

$$\text{minimize } \alpha \cdot \sum_{i \in I} D_i + \beta \cdot V^{\max} \tag{3.15}$$

Constraints

$$\sum_{j \in J_{i,g}} Z_{i,j} = 1 \forall i \in I, g \in G \tag{3.16}$$

$$\sum_{i \in I_j} ZF_{i,j} \leq 1 \forall j \in J \tag{3.17}$$

$$Z_{i,j} \geq ZF_{i,j} \forall i \in I_j \tag{3.18}$$

$$\sum_{i' \notin NC_{i,g}} X_{i',i,g} + \sum_{j \in J_{i,g}} ZF_{i,j} = 1 \forall i \in I, g \in G \tag{3.19}$$

$$\sum_{i' \notin NC_{i,g}} X_{i',i,g} \leq 1 \forall i \in I, g \in G \tag{3.20}$$

$$X_{i,i',g} + X_{i',i,g} + \sum_{j \notin J_{i,g} \cap Ji',g} Z_{i',j} \leq 1 \forall i, i' \in I_g, i' > i, g \in G \tag{3.21}$$

$$Z_{i,j} + 1 - X_{i,i',g} - X_{i',i,g} \geq Z_{i',j} \forall i, i' \in I_g, i' > i, j \in J_{i,g} \cap J_{i',g}, g \in G \tag{3.22}$$

$$\mathcal{M}^1(1 - X_{i,i',g}) + T_{i',g} \geq T_{i,g} + \sum_{j \in J_{i,g}} Z_{i,j} w_{i,g} p_{i,j} \quad \forall g \notin G^{batch}, i \in I, i' \notin NC_{i,g} \tag{3.23}$$

$$\sum_{j \in J_{i,g}} ZF_{i,j} r^j \leq T_{i,g} \forall g \in G, i \in I \tag{3.24}$$

$$r_i \leq T_{i,1} \forall i \in I \tag{3.25}$$

$$\sum_{j \in J} \sum_{b \in B} Q_{i,j,b} = 1 \forall i \in I \tag{3.26}$$

$$\sum_{j \in J} \sum_{b \in B} Q_{i,j,b} S_{j,b} = T_{i,g} \forall i \in I, g \in G^{batch} \tag{3.27}$$

$$T_{i,g-1} + \sum_{j \in J_{i,g-1}} (Z_{i,j}(w_{i,g} p_{i,j}) \leq \sum_j \sum_b Q_{i,j,b} \cdot S_{j,b} \forall i \in I, g \in G^{batch} \tag{3.28}$$

$$T_{i,g} + \sum_{j \in J_{i,g}} (Z_{i,j} \cdot pb_j) \leq T_{i,g'} \forall i \in I, g \in G^{batch}, g' \in NS_{i,g} \tag{3.29}$$

$$T_{i,g} + \sum_{j \in J_{i,g}} Z_{i,j} w_{i,g} p_{i,j} - d_i \leq D_i \forall i \in I, g \in G^{final} \tag{3.30}$$

$$V_t = \sum_i W_{i,t} \forall t \in \{1, \cdots, H\} \tag{3.31}$$

$$W_{i,t} \geq W^1_{i,t} + W^2_{i,t} - 1 \tag{3.32}$$

$$\mathcal{M}^2 W^1_{i,t} \geq t - (T_{i,g} + \sum_{j \in J_{i,g}} (Z_{i,j} pb_j) \forall t \in \{1, \cdots, H\}, g \in G^{batch} \tag{3.33}$$

$$\mathcal{M}^2 W^2_{i,t} \geq T_{i,g} - (t+1) \forall t \in \{1, \cdots, H\}, g \in G^{final} \tag{3.34}$$

$$\text{all variables} \geq 0 \tag{3.35}$$

Each job is processed in each stage exactly once (3.16). There is one job that is the first to be processed on a certain operational machine j (3.17), and a job i can only be processed first on a machine j if the job is assigned to that machine (3.18). Since the successors and predecessors of all orders are tracked, a job i can be processed first on the machine j to which it is assigned, or succeeds another job i' on that same machine (3.19). Furthermore, jobs can only have one direct successor (3.20). Successive jobs i and i' cannot be processed by machines that cannot process them both, but should be processed by the same machine j (3.21) (3.22). These equations were included, since they performed best in the review of [122]. In non-batching stages, all machines are only able to process one job at a time. Therefore, job i' can only start processing on machine j after its predecessor i is finished (3.23). Furthermore, the release time of the machines (3.24) and job (3.25) should be taken into account. Each job i should be assigned to exactly one batch b (3.26), and the batch starting time should be equal to the job timing of each job i in that batch b (3.27). A batch b can start processing after the completion time in the previous stage of the jobs that are in that batch (3.28). The start time of jobs in the post-batching stage, should be later than the batch completion time (3.29). We considered two objectives, the tardiness and the inventory. The tardiness of orders is determined by the completion time of a job i minus the due date of that job (3.30). To determine the inventory at time t, the sum of all jobs that are in inventory at time t is determined (3.31). Using two auxiliary variables, a job is in inventory at time t (3.32) if its completion time in the batching stage is lower than t (3.33), and the start time in the post-batching stage is higher than t (3.34).

In practice, more constraints have to be added to this model. For example working hours of machines and staff may differ (machines can run during the night, while staff is not available). Furthermore, requirements on employee education might be needed for specific job types. These constraints can easily be added to the model, for example using auxiliary variables.

The problem is a Quadratic Problem, due to constraints (3.28) and (3.27). When batch timing decisions are fixed, for example by solving the batching problem, the decision variable $S_{j,b}$ is replaced by a parameter $s_{j,b}$, with fixed start times as derived from the batching model for batch b on machine j. Furthermore, equations (3.28) and (3.27) can be replaced by equations (3.36) and (3.37). This way, the model becomes linear, and thus a MILP.

$$T_{i,g-1} + \sum_{j \in J_{i,g-1}} (Z_{i,j} \cdot (w_{i,g} * p_{i,j}) \leq \sum_{j} \sum_{b} Q_{i,j,b} \cdot s_{j,b} \quad \forall i \in I, g \in G^{batch} \tag{3.36}$$

$$\sum_{j \in J} \sum_{b \in B} Q_{i,j,b} \cdot s_{j,b} = T_{i,g} \quad \forall i \in I, g \in G^{batch} \tag{3.37}$$

CHAPTER 4

Optimization of pathology processes - A case study

4.1 Introduction

Due to varying reasons such as the aging population and the growing awareness of the importance of histopathology for diagnosing disease and assessment of prognosis and therapeutic options, the histopathology laboratory nowadays experiences a growing volume of more complex testing [212]. Furthermore, there is an increasing demand for short turnaround times. At the same time, healthcare costs need to be controlled, and the histopathology workforce is shrinking in many countries [212]. Despite emerging technologies that have facilitated more automation, the histopathology laboratory still operates a sequence of labor intensive, often manual, processes [45, 212, 308]. Therefore, histopathology resources and employees should be deployed as effectively as possible [47]. Moreover, reducing turnaround times in histopathology laboratories is challenging since it is affected by employee workload, the batch-wise way of working, and workflow scheduling [212, 286].

In many hospitals in the Netherlands, tissue processing is still done overnight, due to the long processing times of the larger tissue samples that are part of a batch in the conventional tissue processors. This leads to a one-night delay. Pathologists and histotechnicians adapt their schedules to this overnight tissue processing [308]. This leads unavoidably to unnecessarily high turnaround times (TATs). The implications for the diagnostic workload in the histopathology laboratory are a high workload environment in the morning, and low workload in the afternoon[47]. This requires longer employee recovery times, and comes with detrimental effects, as mentioned in Chapter 3 [102, 202].

In this case study, we apply the mathematical approach developed in Chapter 3, to optimize the process flow in UMC Utrecht's histopathology laboratory by assessing the workload and TAT. This approach offers the possibility to analyze a changing system without physically interfering with the system [327]. The mathematical model was developed using an iterative approach. During the 4 months development, pathologists, residents, and technicians were involved in the design and validation of the model. For example, the categorization of specimens, together with the corresponding process flow of each of the specimen

categories was determined in collaboration with histopathology employees, based on a combination of medical and logistical requirements.

The decomposed MILP model determines the starting times of the tissue processors, and schedules all specimens arriving at the pathology department during the day in the first four stages (see Figure 3.1): grossing, tissue processing, embedding, and sectioning & staining. This results in a tissue processor schedule, together with a schedule of all other activities per specimen. Using the decomposed MILP model, we first analyze the current processes in the histopathology laboratory, whereafter multiple alternative approaches are prospectively assessed. This way, the MILP model not only optimizes the tissue processor schedule, but also assists in scenario evaluation. The outcomes facilitate better decision making on workflow management, resource usage, and TAT, against no additional costs.

The data collection and model setup are described in Section 4.2. Section 4.3 describes the results of the experiments. This chapter concludes with Section 4.4 Furthermore, we show how TAT and workload can be used as effective performance indicators in the histopathology laboratory. Specifically, we show how to prospectively assess laboratory performance using OM/OR methods, and discuss how these methods can assist healthcare decision making. The results of this study are currently deployed for optimizing the workflow in the histopathology laboratory of UMC Utrecht.

4.2 Materials and methods

4.2.1 Laboratory settings

The study is performed at the pathology department of UMC Utrecht. In the pathology department, diagnostics, consisting of histology, immunochemistry, molecular pathology, and cytology is provided for the whole medical center, as well as for other pathology labs, surrounding general practitioners and private clinics. In the histopathology laboratory, tissues of over 30,000 patients are evaluated each year.

Specimens that arrive at the histopathology laboratory go through the predefined sequence of activities, as shown in Figure 3.1. In UMC Utrecht histopathology laboratory in 2013, 69% of the FTE is spent on activities directly related to diagnostics (so-called primary activities). The remainder of time is used for other activities such as educational activities, cleaning activities, and inventory management activities.

Regularly, tissue processing (Step 2 of Figure 3.1) is solely done overnight. However, exceptions were made for priority specimens, which were occasionally processed during the day. This exception allows 100% of the priority specimens that arrived before 10:00 AM in the laboratory to be discussed at the multidisciplinary team meetings (MTM) in the afternoon of the same day.

Usually, every day three residents run the grossing process and 10 technicians

assist in grossing, and perform embedding, sectioning, mounting, staining and case assembly per day. One of these technicians is dedicated to support activities, such as transportation of slides. During the day, the histopathology workforce is supported by a team of pathologists and secretaries. Furthermore, four tissue processing machines are available for tissue processing.

The examination of slides by residents and/or pathologists, the fifth step of Figure 3.1, is excluded from the model. Too many external factors, such as educational and research tasks, influence the behavior of staff in this phase such that no clear examination schedule can be derived.

4.2.2 Performance indicators

We analyze the outcomes of the model in terms of the Intradepartmental TAT (ITAT) performance and inventory distribution over the day. Using chi-squared tests the differences between interventions and the initial situation are evaluated. We assume significant findings when the p-value is ≤ 0.05.

The ITAT is the total time a specimen spends in the histopathology laboratory [257]. We define the ITAT as the number of minutes from arrival of the specimens in the laboratory to transportation of the slides to the pathologist. Furthermore, we consider the Registration TAT (RTAT), which we define as the number of minutes from the moment of transportation until the authorization by the pathologist. The RTAT is not taken into account, as the examination step of the laboratory process is excluded from this research. The total of ITAT and RTAT defines TAT as the total time needed for each specimen to be examined and reported from the moment of arrival in the laboratory.

The inventory level after the embedding stage in number of specimens is a measure for the spread of workload. Note that the number of slides per specimen are not included in this indicator, as employees cannot extract the real workload from a specimen block in the inventory.

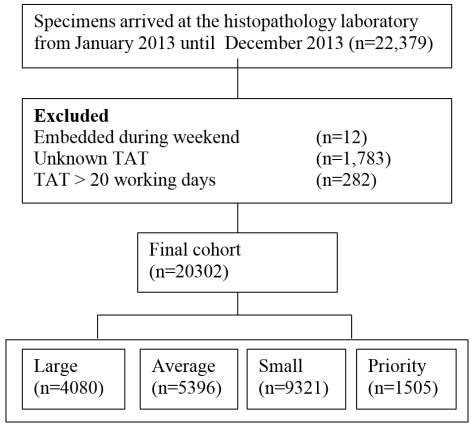
4.2.3 Data collection

The histopathology laboratory of UMC Utrecht has provided real life data for this case study. Work volumes and TAT data are derived from the hospitals Laboratory Management System (LMS). Data on shifts is derived through communication with the lab manager and histopathology staff.

To assess the initial performance, 12 months of historical data are considered consisting of 22,379 specimen samples (biopsies and surgical specimens). This includes all specimens accessioned during regular working hours in 2013 (January 1, 2013 December 31, 2013). Muscle biopsies, which receive many supplementary studies in the testing process, are not included in this database. Exclusion criteria consist of specimens that were embedded during the weekend, or holidays ($n = 12$), or specimens with a TAT of more than one month (> 20 working days) ($n = 282$). For 1,783 specimens TAT cannot be calculated, as the moment of arrival or the moment of sectioning and staining is unknown. The final cohort,

as shown in Figure 4.1 consists thereby of 20,302 specimens, which are included in this study.

Figure 4.1 Flowchart of specimen inclusion



The specimens samples are divided into four job types:

1. *Large specimens*, such as surgical specimens, which need long fixation and are grossed by the residents. The time required for tissue processing is 8 to 12 hours.

2. *Average size specimens*, such as excisional biopsies, which can be grossed by the residents as well as the technicians. In UMC Utrecht, these specimens are often derived from external parties. The time required for tissue processing is 4 hours.

3. *Small specimens*, such as biopsies, which do not need to be grossed, but only to be put into a cassette by the technician. The time required for tissue processing is 3 hours.

4. *Priority specimens*, which are often small sized specimens such as myocardial and breast biopsies. The priority specimens are handled by technicians and processed as soon as possible, preferably with a same-day diagnosis [23]. The time required for tissue processing is 2 hours.

To analyze the performance of possible interventions, we consider 200 different problem instances based the available data. A summary of all input variables and parameter is given in Table 4.1. Each instance includes multiple specimens, which correspond to the jobs in the model. Each job consists of a job type and a number of slides, which vary according to their job type. The number of slides are included as a weight factor to the inventory levels. Furthermore, the moment of arrival of the specimen in the laboratory is included as the release time of the job. The due date of a job is derived from the target turnaround time, as shown in Table 4.2. Furthermore, this table shows the corresponding batch processing time

Table 4.1 Variations in case study variables

Variable	Interval
M_1	1
M_2, B	(4,3), (4,4), (4,5)
M_3	5
F	4
J	37, 66, 95, 105

Table 4.2 TAT targets and processing times per job type

Job type	TAT target	Batch processing time
Job type 1 - large	2 days	720 minutes
Job type 2 - average	8 hours	230 minutes
Job type 3 - small	6 hours	190 minutes
Job type 4 - priority	4 hours	120 minutes

of the job types. The TAT targets per job type, and therefore the corresponding due dates, are set by hospital management, the Dutch government, and external parties, to ensure a timely diagnosis for all patients. The batch processing times are set by the tissue processor manufacturer, academic standards, and laboratory management, as shown in Table 4.2 as well.

The model runs on a laptop personal computer using AIMMS 4.0 mathematical modeling software [36], with MILP solver CPLEX 12.3 [77].

4.2.4 Possible interventions

We selected possible interventions in collaboration with UMC Utrecht pathology employees during multiple group sessions, for example by asking the employees to give input on their perceived optimal laboratory opening hours and shift schedules. The selected interventions are assumed to only influence the ITAT, and not the RTAT. Therefore, if an intervention decreases the ITAT, we believe the overall TAT will be reduced as well. In the remainder of this chapter, we evaluate the following interventions:

Baseline situation (no intervention)

In the baseline situation, tissue processing are only performed during the night, except for one batch of priority specimens. This situation corresponds with the regular way of working of UMC Utrecht histopathology laboratory.

Intervention I: Tissue processing during the day for all specimens

All histopathology employees support the need of tissue processing during the day, to level the workload and shift the peaks from the morning towards the afternoon. The starting times of the various batches during the day are determined by the batching model of Chapter 3.

Intervention II: Staggered shifting

In the initial situation, all staff assigned to a specific activity starts and ends their shifts at the same time. Staggering the shifts may provide a more efficient start in the morning, where a small amount of staff can perform all starting tasks, without other staff waiting for them to be able to begin their job. Furthermore, it ensures that work can be finished in the afternoon, since the late-shift-employees can finalize the cleaning in the end of the day after the other staff went home.

Intervention III: Earlier opening hours

Literature shows that technicians should start embedding directly after the tissue processing batch is finished, to minimize the waiting time in the laboratory [274]. For the night run, the embedding shift should therefore ideally start around 4:00 AM. In the Netherlands these opening hours cannot easily be realized for histopathology staff due to governmental regulations. However, there are opportunities to start one or two hours earlier than the baseline practice.

Interventions I+II: Tissue processing during the day combined with staggered shifts

We hypothesize that more tissue processing during the day leads to a higher workload in the afternoon. This may increase the need for staggered shifts to ensure sufficient workforce in the afternoon. Therefore, we evaluate a combination of Intervention I and Intervention II.

Interventions I+III: Tissue processing during the day combined with earlier opening hours

We hypothesize that when residents start grossing earlier (the direct consequence of earlier opening hours), more tissue can be processed during the day. Therefore, we evaluate a combination of Intervention I and Intervention III.

4.3 Results

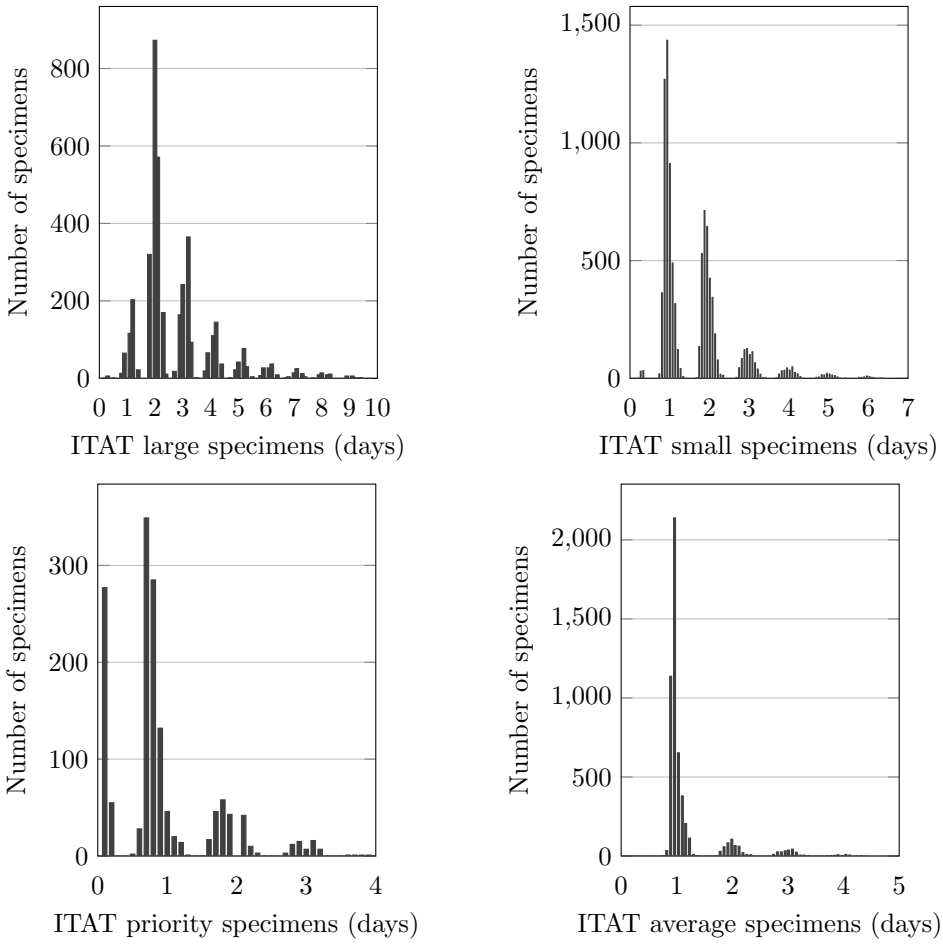
4.3.1 Baseline situation

In total 20,302 specimens are included in this study, from which 4,080 (20%) are large specimens, 5,369 (26%) are average specimens, 9,321 (46%) are small and 1,505 (7%) priority specimens. The mean number of arrivals per day is 80 specimens (range: [34, 122]).

The mean observed ITAT performance is 1.67 days ($\sigma = 1.26$ days), as shown in Table 4.3. The mean observed ITAT varies between the different specimen types, as shown in Table 4.4. Large specimens take 2.79 days ($\sigma = 1.69$ days), small specimens 1.60 days ($\sigma = 1.02$ days), and average specimens 1.16 days ($\sigma = 0.70$ days) on average to be completed. The distribution of ITAT per specimen type is also depicted in Figure 4.2, where every dip indicates an extra night. Figure 4.2 shows that the distribution of ITAT for average specimens and small

specimens differs: small specimens encounter the one-night delay more often, which is shown by a higher mean ITAT.

Figure 4.2 Observed distribution of ITAT in days per specimen type (20,302 specimens)



4.3.2 Validation

To validate the model, output validation and face validation are applied. The output validation compares the ITAT performance of the modeled initial situation, based on a random selection of ten datasets of historical data, to the observed ITAT performance, as shown in Table 4.5 and Table 4.4 respectively. This shows that the observed ITAT performance of the large, average, and small specimens, did not significantly differ from the modeled ITAT performance ($p = 0.883; p = 0.139; p = 0.787$). However, for priority specimens there is a large

Table 4.3 Observed overall turnaround time performance (20,302 specimens)

	Mean	(σ)	Range [min,max]
ITAT [a]	1.67	(1.26)	[0.14, 16.24]
RTAT [b]	2.10	(2.26)	[0.01, 18.00]
TAT [c]	3.77	(2.75)	[0.16, 19.99]

[a] Intradepartmental turnaround time: Time from arrival in the laboratory until transportation towards the pathologist for examination.

[b] Registration turnaround time: Time from transportation towards the pathologist until authorization of the results.

[c] Time from arrival in the laboratory until authorization of the results.

Table 4.4 Observed turnaround time performance per specimen type (20,302 specimens)

		ITAT Mean (σ) in days	RTAT Mean (σ) in days	TAT Mean (σ) in days
Large	(n=4080)	2.79 (1.69)	2.88 (2.69)	5.67 (3.20)
Average	(n=5396)	1.16 (0.70)	1.04 (1.34)	2.20 (1.64)
Small	(n=9321)	1.60 (1.02)	2.48 (2.28)	4.07 (2.52)
Priority	(n=1505)	0.99 (0.78)	1.43 (1.87)	2.43 (2.14)

and significant difference ($p = 0.003$).

For the face validation, pathology employees reviewed the model input and output, and they agreed that the output of the model reflected reality, despite the significant difference in priority specimen performance.

4.3.3 Interventions

As shown in Table 4.5, the modeled ITAT decreases by up to 24% for intervention I ($p < 0.01$), up to 24% for Intervention I+II ($p < 0.01$), and up to 25% for Intervention I+III ($p < 0.01$). Intervention I+III seems to perform best, although ITAT performance is not significantly better compared to Intervention I. Furthermore, the model shows no significant benefits of staggered shifting in terms of turnaround times compared to the baseline situation and to the situation with tissue processing during the day.

In the baseline situation, all jobs are processed in batches during the night, except for a very small amount of priority jobs (1-3 slides per batch), which are processed on fixed moments during the morning. This results in a high workload during the morning, as shown in Figure 4.3 for one representative instance (95 jobs, replication 43). All 95 jobs become available in the morning, which together account for 193 slides. Figure 4.4 shows the spread in workload for the same instance, but now considering Intervention I with 5 batches scheduled during

Table 4.5 Turnaround time performance of the mathematical model for different sample categories

	Large specimens		Average specimens		Small specimens		Priority specimens	
	E(ITAT)[a]	Range[b]	E(ITAT)[a]	Range[b]	E(ITAT)[a]	Range[b]	E(ITAT)[a]	Range[b]
Initial situation	2.87	[2.00, 3.91]	1.49	[1.26, 1.58]	1.51	[1.27, 1.68]	0.25	[0.19, 0.29]
Intervention I [c]	2.82	[1.94, 3.10]	1.00	[0.75, 1.46]	0.92	[0.73, 1.59]	0.24	[0.13, 0.90]
Intervention II [d]	2.89	[2.01, 3.89]	1.54	[1.30, 1.62]	1.57	[1.31, 1.73]	0.29	[0.23, 0.33]
Intervention III [e]	2.85	[1.98, 3.89]	1.52	[1.30, 1.60]	1.54	[1.29, 1.71]	0.25	[0.19, 0.29]
Intervention I+II [cd]	2.82	[1.94, 3.09]	0.95	[0.75, 1.46]	0.94	[0.73, 1.65]	0.25	[0.14, 0.89]
Intervention I+III [ce]	2.80	[1.92, 3.08]	1.01	[0.75, 1.50]	0.90	[0.73, 1.54]	0.20	[0.13, 0.88]

[a] Mean ITAT in days.
[b] Minimum and maximum ITAT in days.
[c] Tissue processing during the day.
[d] Staggered shifts.
[e] Earlier start.

Figure 4.3 Workload performance in the baseline situation

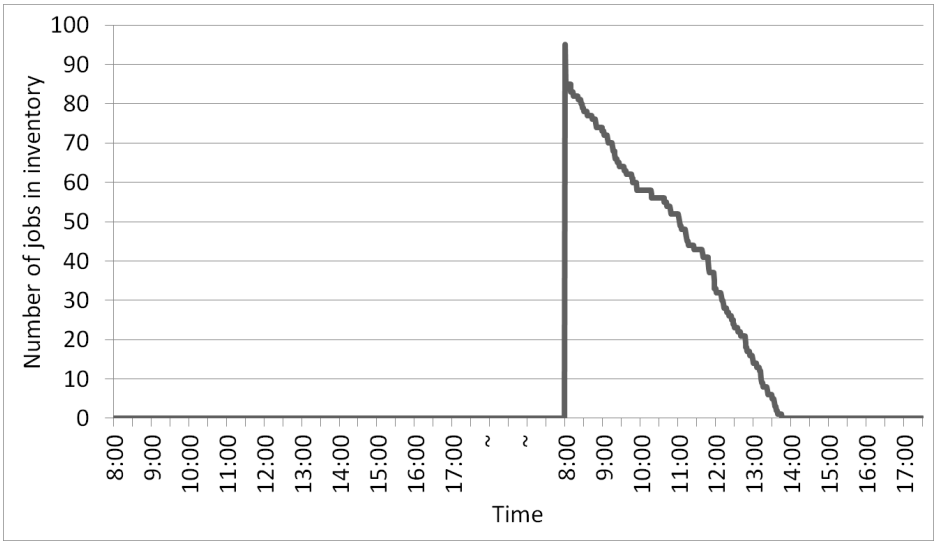
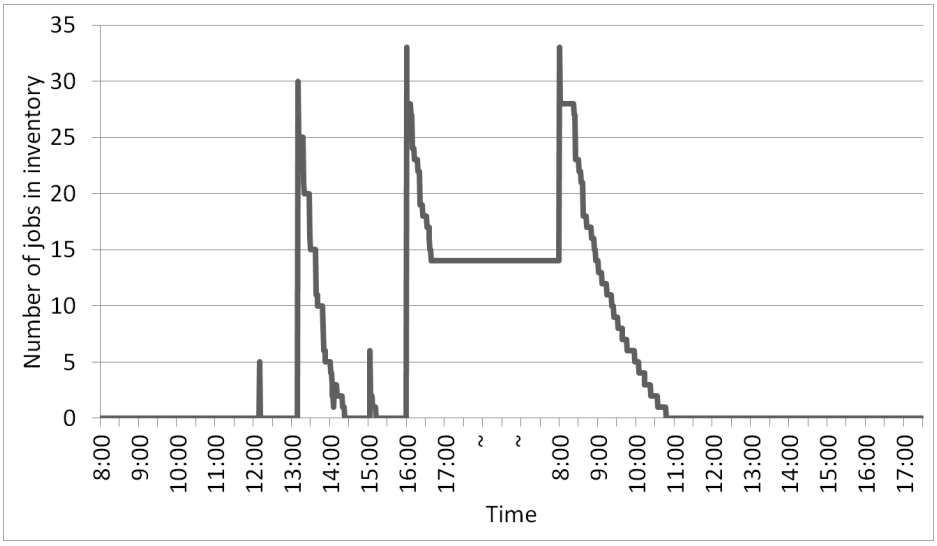


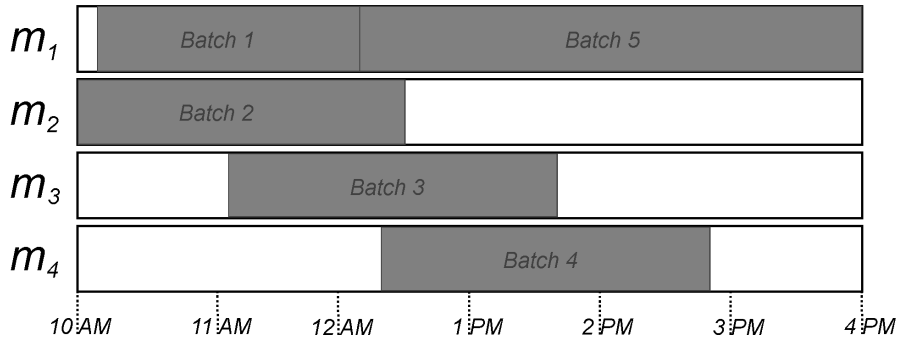
Figure 4.4 Workload performance in Intervention I



the day. Figure 4.5 shows the corresponding batch timing derived from the decomposed MILP model. There is a maximum of 33 jobs in inventory and a maximum of 104 slides. The inventory peak in the morning can be reduced with up to 50%.

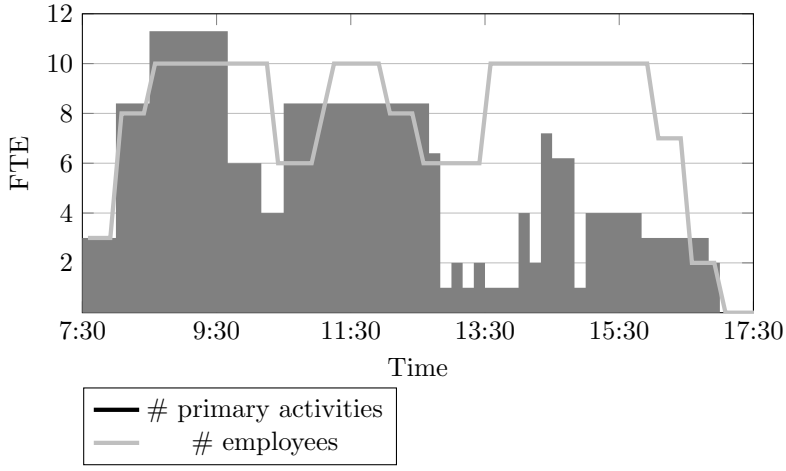
We observe similar results when analyzing the distribution of all primary activities over the day, against the number of available technicians. Figure 4.6 shows that workload peaks occur in the morning, while the afternoon workload

Figure 4.5 Batch timing of Intervention I



is relatively low. At multiple time points the workload in the laboratory exceeds the number of employees available. Introducing tissue processing during the day (Intervention I) results in a more leveled workload, as shown in Figure 4.7. Combined with earlier working hours (Intervention I+III) the morning peaks reduce even more, as shown in Figure 4.8.

Figure 4.6 Workload in the baseline situation – The bars indicate the volume of primary diagnostic activities at a certain time, the gray line indicates the number of employees available at that time in the laboratory



4.4 Conclusions and discussion

The growing workload in the histopathology laboratory while the workforce is shrinking, together with higher demands on TAT, necessitates efficient planning of activities to ensure an equal division of workload over the day and low

Figure 4.7 Workload in the proposed situation with tissue processing during the day (Intervention I) – The bars indicate the volume of primary diagnostic activities at a certain time, the gray line indicates the number of employees available at that time

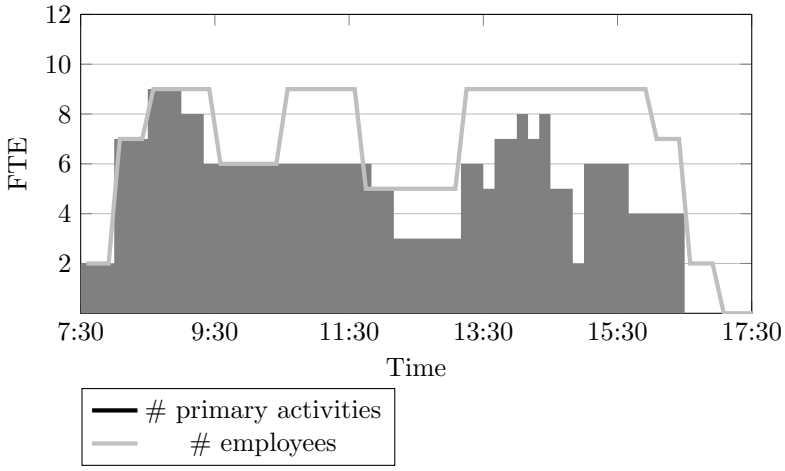
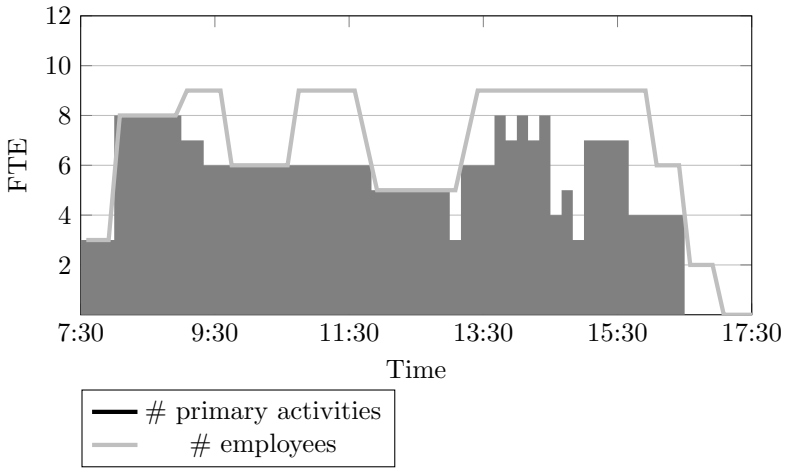


Figure 4.8 Workload in the proposed situation with tissue processing during the day and earlier grossing (Intervention I+III) – The bars indicate the volume of primary diagnostic activities at a certain time, the gray line indicates the number of employees available at that time



turnaround times. Adapting an operations research approach to pathology settings results in valuable insights. The developed model predicts that in UMC Utrecht up to 25% of the initial ITAT can be reduced when tissue processing during the day is implemented and workflow becomes more continuous instead of batch-driven. Furthermore, combining this with grossing small and average sized materials earlier in the morning, the workload will be more leveled throughout the day, with less peak moments in the morning.

In the observed baseline situation, small sized specimens show a higher TAT than average sized specimens, despite their shorter tissue processing time. Most of the average sized specimens are collected from external parties and are therefore received in the end of the day (resulting in a lower ITAT) and analyzed with more priority by a dedicated pathologist (resulting in a lower RTAT), which is the standard way of working in the baseline situation to ensure a rapid diagnosis for these specimens. This can result in prioritization of the average over the small sized specimens.

Surprisingly, in the observations for the baseline situation, the priority specimens show a high ITAT, although we expected that these specimens are processed the same day or within 24 hours. This may be due to a diagnosis registration mismatch. Pathologists deliver the diagnosis at the MTM on time, but often do not succeed in immediately creating the report and registering the diagnosis in the computer system, resulting in a delay of a few hours or to the next morning. The validation showed a large significant difference in observed and modeled ITAT performance for priority specimens, due to this diagnosis registration mismatch. Despite this significant difference, we, together with the pathology employees, consider the model a valid representation.

We observe that the validity of the model highly depends upon the time distribution of specimen arrival. Furthermore, it also depends on the employee behavior. For example, we assumed that all employees work equally fast during every moment of the day and that no overtime is allowed. Therefore, the model sometimes delays the processing of tissue to the next day, while in practice employees work longer, or faster at the end of the day to ensure that all tissue is available for the tissue processing overnight run. Another assumption we made is that all batches have approximately equal loads for the subsequent stages if divided over the day. However, if the arrival pattern of specimens through the day shows high variation, then it can happen that in the grossing stage no samples are available to process at certain times during the day, while at other moments there is a large queue. By including different release time patterns of specimens to the batching problem, the impact of variation can be analyzed and robust solutions can be found. However, including arrival patterns in the model will require intensive computing power.

Due to modeling restrictions of complex human behavior, the RTAT is not included in the decomposed planning model. However, even though we assume that the interventions only impact the ITAT, the interventions may impact the RTAT as well, and therefore increasing or decreasing the TAT. Further research should be executed to collect empirical evidence to evaluate the actual impact on TAT.

The model shows that through earlier starting grossing shifts in combination with tissue processing during the day, more tissue can be grossed before the start of tissue processing batches, which allows for more tissue to be finished the same day. We notice that using the model, the small sized specimens outperformed the average sized specimens when tissue processing during the day is allowed, even though the average sized specimens have more strict turnaround time targets.

This is a direct consequence of the lower processing times for tissue processing batches of small sized specimens and the moment of arrival of average sized specimens.

The model shows no significant benefits of staggered shifting in terms of ITAT compared to the baseline situation and the situation with tissue processing during the day. A possible explanation might be that the current number of specimens is not large enough to make this a beneficial intervention.

Reasoning from the principle that workload and staff resources need to be balanced, a more balanced distribution of work over the day can be derived by carefully planning the amount of work over the day, by carefully planning the required staff resources over the day, or by optimizing a combination of both. Changing the number of staff resources can result in balanced solutions, for example by adapting the schedules of the technicians to the actual amount of work. By adding more short shifts in the morning and the late afternoon (for example from 9:00 to 12:00, and 16:00 to 18:00), the peaks in the workload can be covered. However, due to regulations in the Netherlands, there are limitations to flexibility to create early and late shifts, and working before and after regular shifts may require extra financial compensation. Visualization of the amount of work corresponding with each of the interventions shows that the tasks as scheduled in the baseline situation are shifted towards other moments during the day when introducing tissue processing during the day. For example, embedding and sectioning shift from the morning towards the afternoon. Furthermore, especially on busy days, tissue processing during the day results in a more equally distributed workload. A combination of changing the amount of work and the number of resources results in the best solutions. This more continuous workflow is likely to be preferred over the traditional batch-mode workflow [210], as is also well known from other disciplines [323].

Given the results of the decomposed model, the workflow of the histopathology laboratory of UMC Utrecht has been redesigned. The staff decided to implement interventions I and III, in order to better spread the workload over the day, and assure lower TAT for all patients. As shown in Figure 4.8, even in the proposed situation there still is a gap in primary activities in the afternoon. Ideally, a batch from the tissue processor should finish at this particular moment. However, due to the low amount of tissue eligible for tissue processing during the day, this option is currently not feasible in UMC Utrecht.

Besides the fact that Intervention II does not show any improvements compared to the initial situation in the histopathology laboratory, it is not adopted because of practical reasons as well. Every time a new technician starts his or her shift, they will inevitably loose some time on social interaction with their colleagues, which is expected to increase with staggered shifts.

Further research should be executed to investigate whether there is a relationship between the quality of work and the more leveled workload. When employees experience less stress during their work, the quality of their work is likely to improve [102]. This can be evaluated by the number of errors and rework, for example in the number of slides that have to be reprocessed and mix-ups.

Furthermore, stress levels can be measured before and after interventions.

Besides turnaround time in the histopathology laboratory, as measured in this study, the hospitals specimen turnaround time is often measured from the time the biopsy is taken until the time the report is sent to the clinician. The model allows for extensions, for example by including transportation from the biopsy room to the laboratory, and the biopsy itself. However, during this research, it was hard to include the evaluation of specimens by the pathologists, since many factors influence this process. Therefore, the process of examination has to become more standardized before it can be reliably modeled.

Same day reporting becomes more important in pathology [213]. To facilitate same day reporting, dedicated resources are often assigned to specific specimen categories, which are therefore prioritized. This study shows that in UMC Utrecht histopathology laboratory specific specimen types are (unintentionally) prioritized over the remaining specimens. This may cause other specimen types to underperform, since Vanberkel et al. [297] showed that dedicating resources and prioritizing categories increases the TAT of the remaining care. Therefore, further research should be done to investigate the presence and implications of prioritization of specific specimen processing over the remaining care.

In conclusion, the decomposed MILP model is valuable in designing the histopathology processes, by assessing the impact of optional interventions and optimizing scheduling decisions. The proposed interventions are implemented in the histopathology laboratory of UMC Utrecht. Further research is needed to collect empirical evidence to evaluate their actual impact on ITAT, TAT, quality of work, and employee stress levels.

PART 3

outpatient clinic

Why Wait?
Organizing Integrated Processes in Cancer Care

Chapter 5

A.G. Leeftink, M.G. Martinez, E. Sisikoglu Sir, E.W. Hans, M.Y. Sir, and K.S. Pasupathy. Scheduling interval optimization in healthcare clinics considering no-shows and cancellations. *Submitted*.

Chapter 6

A.G. Leeftink, I.M.H. Vliegen, and E.W. Hans. Stochastic integer programming for multi-disciplinary outpatient clinic planning. *Health Care Management Science*, `https://doi.org/10.1007/s10729-017-9422-6`, 2017.

CHAPTER 5

Scheduling window under no-shows and cancellations

5.1 Introduction

No-shows and cancellations for outpatient clinic appointments result in adverse outcomes, both for the clinics as well as for their patients. In order to mitigate the effects of no-shows and cancellations, this chapter analyzes the no-show and cancellation behavior of outpatient clinic patients, as well as a queuing approach to incorporate this behavior in the design of these clinics.

5.1.1 Effects of no-show and cancellation rate

Ever since the increasing focus on efficient healthcare operations, clinics started to evaluate their no-show and cancellation rates. No-shows and cancellations result among others in reduced productivity and efficiency for hospitals [81], financial impact through reduced revenue and idle resources [9, 26, 90, 208, 221], reduced learning opportunities for residents [123], and the waste of valuable resources, which could have been used to serve other patients, as canceled slots cannot always be filled with new arrivals, and missed appointments can only be filled by walk-in patients. Furthermore, no-shows and cancellations increase the waiting lists, by reducing the number of appointments available. Therefore, it reduces patient access to care [37, 81, 90, 165, 221]. This might affect the continuity and quality of care for patients [26, 38, 140, 165]. For example diabetic and hypertension patients who frequently miss appointments are more likely to have worse health outcomes [192, 220]. Furthermore, the reduced patient access to care can cause a vicious cycle, with longer waiting lists increasing the non-attendance rates, which in turn increases the waiting times again [136].

5.1.2 Cancellation behavior

Although appointment attendance behavior has been studied for over half a century [221], the high volume of recent medical research on this topic shows the problem is still present in healthcare systems. However, most of this literature only distinguishes between no-shows and shows, and excludes cancellations as a

specific category from the analysis [221]. In such studies, cancellations are either included as no-shows [104, 116], included as shows [126], or excluded from the analysis all together [173, 180]. Only a few recent studies have analyzed no-shows and cancellations as two separate conditions [132, 221, 228, 276], despite the different behavior of patient cancellations compared to patient no-shows [132].. It is important to analyze patient cancellation behavior in isolation, as canceled appointments give opportunities to reallocate capacity [132, 221], and therefore to increase the clinic's utilization, and to increase the number of patients that gets access to the clinic.

For clinics it might be challenging to fill appointment slots after last-minute cancellations, resulting in an idle resource, which has a similar effect as a no-show. Similar reasoning holds for patients that want to reschedule their appointment at late notice. To be able to assess this opportunity loss of canceled patients, it is important to not only take the amount, but also the timing of cancellations into account. We define the cancellation interval as the number of business days from the creation of an appointment to the date the appointment is canceled. As an example, Chariatte et al. [65] stated that in their healthcare institution there might be a peak in last-minute cancellations, by patients that want to avoid a payment for a missed appointment. To the best of the authors' knowledge, data on the cancellation interval is not reported before in the literature.

5.1.3 Predictors of no-shows and cancellations

Most research focuses on identifying predictors for no-shows and cancellations, such that targeted interventions can be developed to mitigate the effects [37, 126]. The literature is divided about the impact of demographic characteristics (such as gender, marital status, race, and social class) on the no-show rate and cancellation rate. Where some showed predictive power, others did not find a relation, as these factors are practice and context specific [26, 37]. The show-up history of a patient is related to the no-show probability [65, 317], and the patient's age and clinic and visit type are often found as predictors for no-show and cancellation behavior [10, 26, 37, 81, 126, 173, 228, 255]. Finally, scheduling characteristics are found to be predictors of the no-show and cancellation rates. Multiple studies show significant relations between a day of the week, time of the appointment, scheduling interval and the no-show and cancellation rates [234, 289].

The relationship between the scheduling interval and the no-show and cancellation rates is well-studied. We define the scheduling interval, also referred to in the literature as lead time, planning horizon, appointment age, or appointment interval, as the number of business days from the creation of the appointment to the date the appointment is scheduled for. Benjamin-Bauman et al. [29] and Festinger et al. [98] both analyzed the impact of the scheduling interval on the no-show rate using an experimental design. They randomly assigned patients requests to an appointment slot with a specific scheduling interval, and monitored among others the no-show behavior. They found that patients had higher probabilities of showing up when their scheduling interval was shorter. In predictive

studies, Norris et al. [221] and Bean and Talaga [26] found that the scheduling interval is the most significant predictor of patient non-attendance, both for no-show and cancellation rates. Whittle et al. [321] found a modest effect of the scheduling interval on no-shows, as for large scheduling interval the no-show rate stabilized. Furthermore, they found a highly significant effect of the scheduling interval on cancellations. Partin et al. [228] found the scheduling interval to be a predictor of both no-shows and cancellations as well. Besides many studies that found a significant relation with the scheduling interval and cancellations and/or no-shows [37, 81, 104, 126, 173, 191], some studies did not find such a relation between the scheduling interval and no-show rate [59, 317]. Concluding, patients that have a longer scheduling interval tend to have a higher probability of no-show and cancellation. However, when the scheduling interval becomes very long, these effects may fade out [26, 321].

5.1.4 Strategies to impact patient no-show and cancellation behavior

To mitigate the effects of no-shows and cancellations, clinics can try to influence patient behavior or modify their scheduling strategy [79].

Patient behavior can be influenced by education, reminders, and financial rewards or penalties.

Education can be provided through entrance and exit interviews [123], and through orientation and reminder letters [272].

Reminders can be used for a targeted patient population with a high non-attendance rate [276]. Examples of reminders are SMS reminders [38, 90, 124, 165, 237, 289], email reminders [180], postal reminders [140], and telephone reminders by staff or automated [140, 227, 237, 276]. From the various reminder types, SMS reminders are considered the most cost-effective [237]. Education and reminders are in general considered as an effective method to improve the no-show rate by 10% to 50%, although, in some specific settings, their effectiveness could not be proven (i.e.,[72]). However, there is an inverse effect of reminders on the cancellation rate [165, 166, 276, 289]. A possible reason for this could be that when patients are reminded of their appointment, a portion of the patients that would have resulted in a no-show since they forgot about the appointment, now cancel their appointment. The timing of reminders might influence the attendance rate. Henderson [140] found that telephone and postal reminders are most effective when sent one day before the actual appointment, which suggests there is a time-dependency to no-show behavior. In a literature review on SMS reminders, Boksmati et al. [38] found no study that analyzed the relationship between the timing of the reminders and attendance rates, which is therefore an area of further research.

Financial rewards or penalties do not result in higher attendance rates, according to Chariatte et al. [64]. They might result in higher last-minute cancellations against lower no-shows, as more people will take the effort to cancel an appointment instead of not showing for their appointment [65]. Other penalties

are for example a (temporary) termination of access to a hospital after a no-show. These strategies to influence patient behavior are in general less favorable, as they might impose a financial burden to the access of care [79].

5.1.5 Scheduling strategies minimizing the effect of no-shows and cancellations

Besides strategies to impact patient behavior, scheduling strategies can be adopted. Scheduling strategies that aim to minimize the adverse effects of no-shows and cancellations include overbooking, open access scheduling, panel sizing, and minimizing the scheduling interval.

When *overbooking* is allowed, additional patients are booked to timeslots with a high probability of becoming idle, or booked in overtime, based on the probability that patients cancel or miss their appointment [144, 167, 168, 185, 255, 324]. This way, the probability of resource idle time is minimized, and patients can get earlier access. However, overbooking may increase direct waiting times for patients that show up for their appointment, which could result in reduced patient satisfaction and lower attendance rates on the long term [79].

Open access scheduling (also known as walk-in scheduling) schedules patients that require an appointment the same day, or allows patients to be seen at a walk-in basis [215, 251, 261]. Since the scheduling interval is (close to) zero in this situation, the impact of cancellations and no-shows is small. However, high fluctuations in daily demand may result in idle and overtime. The hybrid policy, in which patients can both schedule an appointment or walk-in, allows using the idle time caused by no-shows to serve walk-in patients. However, Moore et al. [208] showed that using a walk-in visit to cover idle time, does not lead to complete financial recovery, even when it leads to full utilization.

A third scheduling strategy is *panel sizing*. Panel sizing limits the number of patients allowed in the patient panel, which includes all patients that can potentially use the service of the provider. This way, the waiting list can never explode, as the number of patients that can get admitted is controlled [116]. Through the waiting list length, the number of no-shows and cancellations is controlled as well, as patients that are waiting longer have a higher no-show and cancellation probability. However, most outpatient clinics cannot limit their patient population, which makes this strategy especially valuable for the primary care setting.

Liu [184] recently explored the idea of *limiting the maximum scheduling interval* instead of the panel size. Using this scheduling window, one can control the waiting list as well, and thus the number of no-shows and cancellations. However, rejecting all patients that require an appointment outside the scheduling window, might result in patient loss and under-utilization of the system [321]. Therefore, Liu [184] developed an M/M/1/K queuing model, which penalizes the patient loss, and considered a small revenue for empty slots, both due to under-utilization and no-shows.

5.1.6 Research aim and outline

Concluding, medical literature starts to recognize the need to include cancellations into non-attendance analyses, as canceled appointments give opportunities to reallocate capacity [221]. However, scheduling strategies do not take cancellations into account. Furthermore, as we hypothesize that both no-shows and cancellations depend on the scheduling interval, time-dependent no-shows and cancellations should be taken into account, whereas most literature assumes fixed cancellation and no-show rates, which is thus independent of the scheduling interval [2]. This leads to the following research question:

How to use scheduling interval-dependent no-show and cancellation rates in a scheduling approach to improve a clinic's appointment system design?

To answer this question, we first perform a data analysis based on real life data of two major healthcare institutions from the USA and the Netherlands, as well as data from the literature, to analyze the time-dependent behavior of no-shows and cancellations and to show practical relevance. We not only statistically show a monotonic increase of the no-show and cancellation rates depending on time, similar to most medical literature, but also fit a distribution to this behavior. Furthermore, we analyze the cancellation behavior in more depth, to assess not only whether a cancellation happens within the scheduling interval, but also when this cancellation happens, as the timing of a cancellation impacts the possibility of reusing the canceled time-slot. The analyses show that cancellations are more frequent than no-shows, and may seriously impact a clinic's operational efficiency. Therefore, not only time-dependent no-show behavior, but also cancellation behavior should be taken into account when designing appointment systems.

After the data analysis, which serves as input for the scheduling approach, we build on the scheduling approach of Liu [184] to answer the research question. This approach allows for implementation in outpatient clinical practice, and includes the time-dependency of no-shows and cancellations. We use an M/M/1/K queue, similar to the work of Liu [184] and Green and Savin [116]. However, our work differs from these studies as they do not consider cancellations, for they are excluded [184] or included as no-shows [116]. To include the time-dependent cancellation behavior, we use a queuing model with reneging. Using this model, we optimize the scheduling interval, such that cancellations and no-shows are minimized, but provider utilization and patient acceptance are maximized. We show that the optimal scheduling window highly depends on the time-dependent no-show and cancellation rates, and may therefore vary for various clinic types. Furthermore, we compare the impact of the scheduling window on no-shows and cancellations, to enable healthcare managers to determine where to prioritize their interventions. This shows that for high now-show and cancellation/high demand clinics, it is beneficial to keep the scheduling window as small as possible, whereas for low demand/low no-show and cancellation clinics, the scheduling window can be extended.

Our contribution is threefold: (1) We analyze the time-dependency of no-shows and cancellations, together with the timing of cancellations. (2) We compare no-show and cancellation behavior from two health systems in the USA and the Netherlands. (3) We develop a mathematical model to determine the optimal scheduling interval in which we are the first to take time-dependent no-shows and cancellations into account.

The remainder of this chapter is organized as follows. Section 5.2 analyzes the time-dependent behavior of no-shows and cancellations, to show the practical relevance of our research question. Section 5.3 presents the model and gives some structural properties. Sections 5.4 and 5.5 present the simulation model and experiments, and Section 5.6 gives the conclusions and discussion.

5.2 Practical relevance

To design an appointment system that incorporates no-show and cancellation behavior in Section 5.3, we need to get insight into this behavior. Based on the literature analysis of Section 5.1, we hypothesize the no-shows and cancellation rates to depend on the scheduling interval. To show the practical need to include this time-dependent behavior in the design of appointment systems in healthcare, this section presents applications from a large medical center in the USA and a large medical center in the Netherlands. The data collection is described in Section 5.2.1. Next, we present the no-show and cancellation outcomes in Section 5.2.2. We summarize our results in Section 5.2.3.

5.2.1 Data sources

We included retrospective appointment scheduling data from two hospitals, namely Mayo Clinic in Rochester, MN, USA, and University Medical Center Utrecht in the Netherlands. These institutions will be arbitrarily referred to as Institution 1 and Institution 2 in the remainder of the chapter. Data of about 32,000 appointments was extracted from the hospital information system of Institution 1, and data of about 42,000 appointments was extracted from the hospital information system of Institution 2.

The data set of Institution 1 consists of almost 3 years of data (2014/01/01-2016/10/31), and includes all appointments that were scheduled during this time interval in four clinics. The data set of Institution 2 consists of 2 years of data (2015/01/01 - 2016/12/31), and includes all appointments that were scheduled in one clinic during this time interval. The clinics serve, among others, neurology and otorhinolaryngology patients, using an appointment system with fixed slot sizes. No walk-in patients are served in these clinics. The obtained data fields are summarized in Table 5.1. Using the data, we derive three additional data inputs per appointment. The first field is the disposition status, where appointments are clustered in three categories, *Seen*, *Canceled*, and *No-show*. Each appointment where a patient showed up for his or her appointment is classified as Seen. When

Table 5.1 Data inputs

Field name	Explanation
appointment number	Unique appointment number
appointment date	Date on which the appointment is scheduled
appointment create date	Date on which the appointment is created
appointment disposition date	Date on which the appointment status is changed
disposition status	Status of the appointment: seen, canceled, or no-show
disposition reason	Reason for a change in the appointment status, empty if the status is seen
scheduling interval	Time between the appointment create date and the appointment date
cancellation interval	Time between the appointment create date and the appointment disposition date

an appointment is canceled or rescheduled more than 24 hours in advance, it is classified as Canceled. Patients who are not present at their appointment without any notice, who are hospitalized, who are denied for service, and appointments that are canceled or rescheduled within 24 hours of the actual appointment, are registered as a No-show. The second field is the scheduling interval, which is the number of business days from the creation of the appointment to the date the appointment is scheduled for. This includes the scheduled date, but excludes the create date. For example if an appointment is created on Thursday, and scheduled for the following Tuesday, the scheduling interval is 3 days. Note that same day appointments have a scheduling interval of 0 days. The third field is the cancellation interval, which is the number of working days from the creation of an appointment to the date the appointment is canceled. This includes the create date, but excludes the cancellation date. For example if an appointment is created on Thursday, and scheduled for the following Tuesday, but canceled on Monday, the cancellation interval is 2 days. This parameter is only determined for canceled appointments, using the appointment create date and disposition date. All data is represented in business days, and all data fields are summarized in Table 5.1. We limited our study to face-to-face appointments, with a nurse practitioner or clinician. Furthermore, as both hospitals also have education and research tasks, we only included care related appointments.

Sewitch and Hosseina [275] recently raised some important questions regarding the data collection on canceled and missed appointments. They notice that misclassification can occur, since a canceled appointment can be rescheduled and end up in a no-show. However, as we consider reschedules as being new arrivals, no such misclassification can occur. Furthermore, they were concerned that the reason for cancellation might have influenced the data, as a cancellation can occur by patient-initiation, but also be initiated by the clinic. As clinic initiated cancellations reflect system behavior, these cancellations are likely to behave dif-

ferently [275]. Therefore, Whittle et al. [321] and Blæhr et al. [37] performed two analyses for both patient initiated and clinic initiated cancellations. Both studies found significant relations observed similar cancellation rate behavior for patient and clinic initiated cancellation rates. Furthermore, Foreman and Hanna [101] analyzed the impact of the scheduling interval on attendance rates, and found that the impact is independent of the reasons for non-attendance.

5.2.2 No-show and cancellation rates

This section presents the no-show and cancellation rates based on data from the literature, based on the datasets of the two institutions. The four clinics from Institution 1 from are part of the same department and showed similar no-show and cancellation trends. Therefore, in what follows, we present aggregated results for these four clinics. To analyze the no-show and cancellation rates, we perform several statistical tests, with the no-show and cancellation rates as dependent variables, and the scheduling interval and cancellation interval as independent variables. Spearman's rho correlation coefficients are calculated to assess whether there is a monotonic relationship between appointment disposition and the scheduling interval. To evaluate whether the cancellation-motivation impacts our hospital data, we perform a subgroup analysis for patient-initiated and clinic-initiated cancellations. To analyze the timing of cancellations, Spearman's rho correlation coefficients are calculated to assess whether there is a monotonic decreasing relationship between appointment disposition and the scheduling interval. Furthermore, we perform a subgroup analysis for patients with various scheduling intervals, to determine the timing of cancellations. We use IBM SPSS Statistics 22 for Windows for all statistical analyses.

Real life data based no-show and cancellation rates

For Institution 1, the no-show rate slightly increases from 10.3% for appointments that are scheduled the next day to 16.3% for appointments that are scheduled 50 days in advance (see Figure 5.1). A weak positive monotonic correlation is found between the daily scheduling interval and the no-show rate (Spearman's ρ = 0.344, n=61, p=0.007).

The cancellation rate increases from 12.3% for appointments that are scheduled the next day to 42.0% for appointments that are scheduled 50 days in advance (see Figure 5.1). A strong positive monotonic correlation is found between the daily scheduling interval and the cancellation rate (Spearman's ρ = 0.741, n=61, p<0.001).

For Institution 2, the no-show rate slightly increases from 9.1% for next day appointments to 11.0% for appointments that were scheduled 50 days in advance. A weak positive monotonic correlation is found between the daily scheduling interval and the no-show rate (Spearman's ρ = 0.230, n=61, p=0.075).

The cancellation rate increases from 8.9% for next day appointments to 37.7% for appointments that were scheduled 50 days in advance. A strong positive

monotonic correlation is found between the daily scheduling interval and the cancellation rate (Spearman's $\rho = 0.877$, n=61, p<0.001).

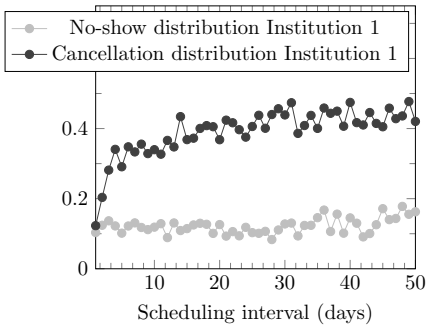


Figure 5.1 No-show and cancellation distributions per scheduling interval in days for Institution 1

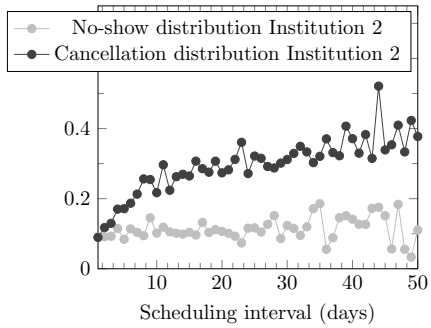


Figure 5.2 No-show and cancellation distributions per scheduling interval in days for Institution 2

Section 5.1 showed two cancellation categories are distinguished: Immediate cancellations and late cancellations. As it is hard to distinguish this behavior for appointments with a short scheduling interval, the first timeslots in Figures 1 and 2 may have impacted the significant correlation. Therefore, we analyzed the cancellation correlation excluding the first 3 data points in the time line, and found a similar strong positive significant relation for both cancellation rates (=0.699, p¡0.001, n=58; =0.857, p¡0.001, n=58), which shows this impact is negligible.

Approximation of exponential distribution

Figure 5.1 and Figure 5.2 show the no-show and cancellation rates are increasing in the scheduling interval. This is in line with the findings of Green and Savin [116], Liu [184]. Green and Savin [116] propose the following no-show rate:

$$\nu_j = \nu_{\max} - (\nu_{\max} - \nu_0) \exp^{\lfloor -j/\mu \rfloor / C},$$

where $\nu_{\max}$ reflects the maximum observed no-show rate, ν_0 the minimum observed no-show rate, and C is a scaling parameter. As μ is the service rate and j the number of timeslots, j/μ is the number of days in the scheduling interval. Similar reasoning holds for the cancellation rate:

$$\chi_j = \chi_{\max} - (\chi_{\max} - \chi_0) \exp^{\lfloor -j/\mu \rfloor / C}.$$

We find the best-fit values for the parameters by minimizing the sum of the mean squared errors between the observed no-show and cancellation rates and the expected rates from the functions. This way we find a no-show rate and cancellation rate for each institution, which are displayed in Table 5.2.

Table 5.2 Parameter settings for no-show and cancellation rates per scheduling interval in days

	$\nu_{\max}$[a]	ν_0[a]	C
No-show rate Institution 1	0.137	0.083	13
No-show rate Institution 2	0.181	0.096	97
Cancellation rate Institution 1	0.457	0.000	5
Cancellation rate Institution 2	0.311	0.000	7

[a] For cancellation rates this reflects χ.

Cancellation timing

Besides the no-show and cancellation rates, we are also interested in the timing of the cancellations. As no timing behavior is reported in the literature, we hypothesize that patients cancel their appointments both early and late in the scheduling interval, as they realize right after scheduling the appointment that a date is not convenient, or realize when the appointment is coming closer that for example other commitments are more important than this appointment. As we expect this behavior to be more distinct for patients with larger scheduling intervals, Figure 5.3 shows the cancellation timing behavior for Institution 1 for various subgroups based on increased scheduling intervals (similar results for Institution 2 not shown). We normalized the scheduling intervals on the interval [0, 1], with 0 being the date on which the appointment is created, and 1 the appointment date. As Figure 5.3 shows, the probability of a cancellation happening in the scheduling interval indeed follows a bimodal distribution, with a peak right after the create date of the appointment, and right before the appointment date. The frequency plots for appointments scheduled within 5 days are not shown, as the bimodal behavior is especially visible for cancellations with larger scheduling intervals. For small scheduling intervals both peaks merge into one peak.

Initiation of cancellations

Institution 1 initiated 11% of the total cancellations, which shows the majority of cancellations is patient initiated. Institution 2 initiated 42% of its cancellations. The main reason for clinic initiated cancellations are scheduling errors. Both patient and clinic initiated cancellations show similar significant monotonic increasing behavior. Furthermore, the timing of cancellations in both groups shows a similar pattern as well, except for a small increase in early cancellations in the clinic initiated cancellation group. A comparison of the cancellation rates for the canceled appointments with a scheduling interval of 1-2 months is shown in Figure 5.4 for Institution 1. Reasons for the clinic to cancel the appointment are related to scheduling errors and unexpected changes in provider calendars due to for example illness.

Figure 5.3 The probability of the timing of a cancellation for a given scheduling interval

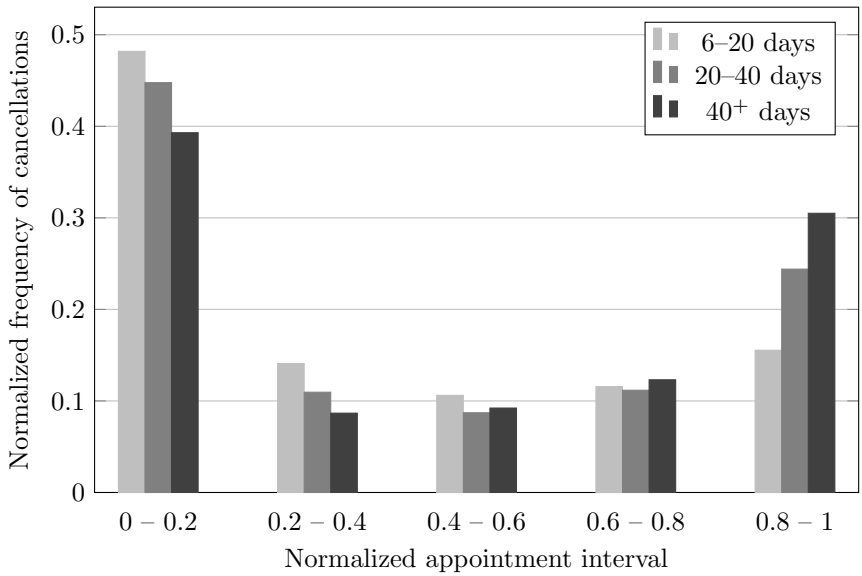
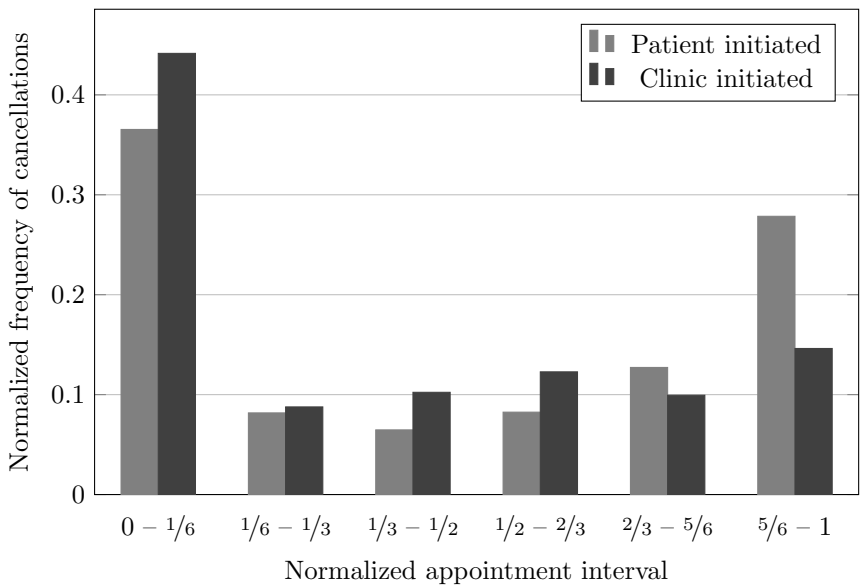


Figure 5.4 The probability of the timing of a cancellation per initiation subgroup for appointments with a scheduling interval of more than 2 months



5.2.3 Summary of the results

This section analyzed the no-show and cancellation behavior of two healthcare systems. We analyzed both a USA and a EU based clinic, and we conclude that no-show and cancellation behavior is similar for the various health systems, as monotonic increasing rates are observed, as well as bimodal cancellation timing behavior.

This is the first study to analyze the timing of cancellations. We observe bimodal behavior, with two cancellation peaks, right after the moment that the appointment is scheduled, and right before the actual appointment time. This is an important observation, as slots of appointments canceled in the first peak can be reassigned with a high probability to new patients. However, slots of appointments canceled in the second peak are less likely to be reassigned. This effect has to be taken into account in the design of appointment systems.

A comparison of the obtained no-show and cancellation rates shows that the no-show rate converges faster than the cancellation rate. This is in line with the literature [321]. Therefore, we expect that reducing high scheduling intervals does not influence the no-show behavior of patients too much, but highly impacts the cancellation rates.

Concluding, we observe scheduling interval dependent no-show and cancellation rates for several data sources from the literature and from USA and EU practice. As this impacts the possible performance of an appointment system in clinics, these systems need to be designed and optimized taking the time-dependent behavior into account.

5.3 Queueing model

Given the time dependency of no-show and cancellation rates and their negative impacts on clinic efficiency and provider productivity, we developed a queueing model to determine the scheduling interval needed in order to minimize the system inefficiencies introduced by these rates. This section presents the mathematical formulation of the problem.

We consider a single-server queuing system with no-shows, reneging in the queue, and balking to evaluate the optimal scheduling interval. Patients are served on a First Come First Serve (FCFS) basis, and due to the finite capacity of the appointment system, patients that arrive with $K-1$ patients in the queue will leave. Cancellations are patients who randomly leave the queue before their appointment. Furthermore, the system encounters no-shows. When a patient does not show-up for an appointment, the servers will be empty for the entire service time of this patient. No overtime, and no preemption of service is allowed, and similar to Liu [184] and Green and Savin [116] we assume exponential service times.

The appointment system can be accurately represented by a multi-server queuing model with deterministic service times, in which the number of servers

represent the number of appointment slots available on one day. However, approximations are needed to analyze this type of system since multi-server queues are analytically intractable. In this study, we assume that patients are offered one appointment slot on a FCFS basis, and service times of appointments are exponentially distributed with mean time μ. Under these assumptions, the system can be modeled as a single-server queuing model with capacity K, i.e., an M/M/1/K queue with μ appointment slots provided in a unit of time [184].

Our objective is to optimize the scheduling interval, which in units of time is equal to $\lfloor K/\mu \rfloor$. A solution is considered optimal when the expected system revenue is maximized. This revenue is a combination of an added value of serving patients and a penalty for blocking patients. The system reward for serving patients increases with small values of the scheduling interval because more appointment slots could be filled. Limiting the scheduling interval on the other hand, impacts access for patients, thus increasing the system penalty for blocking patients. This way, a trade-off will be derived.

In the remainder of this section we derive the long-run behavior of the system under patient no-shows, cancellations, and balking. We assume patients arrive from an infinite source according to a Poisson distribution with rate λ. Patients are served by a single server with exponential service rate μ. Patients do not enter the queue if they encounter a full queue at their arrival, i.e., if the number of patients in the queue is $K-1$. Each patient is rejected at an opportunity cost θ_B. If the queue is not full, i.e., the number of patients in the queue is less than $K-1$, a patient always enters the queue.

Each patient in the queue who cancels his/her appointment generates a cost θ_C since this patient is lost. A patient waits a random amount of time before canceling, which is assumed to have a negative-exponential distribution with constant rate α. The rate α represents the average number of cancellations of the system per unit of time, which is assumed to be independent of the queue capacity K. Consequently, the long-run probability that any one of the j patients scheduled in the system may cancel his/her appointment appointment is equal to $c_{j+1} = j\alpha$, $j = 0, \ldots, K-1$ [13]. We have that the cancellation probability of the system is strictly monotonic with respect to the length of the queue $0 = c_1 < c_2 < \cdots < c_K$. Converting to units of time, the cancellation probabilities have a non-decreasing behavior as time increases: $0 = c_{\lfloor 1/\mu \rfloor} < c_{\lfloor 2/\mu \rfloor} < \cdots < c_{\lfloor K/\mu \rfloor}$.

Using the idealized cancellation rate, we can obtain tractable formulations of the steady-state probabilities of the queuing system. Including the time-dependency of the cancellations shown in Section 5.2 requires more involved models, which are left to be studied in future research. Section 5.4 presents a simulation model which uses empirical cancellation timing distributions derived from historical appointment data, to determine the impact of this assumption within the context of a real appointment scheduling system.

Each patient that enters the queue and does not cancel before service, has a probability of missing the appointment. The probability that a new arrival will be a no-show when upon arrival there are j patients scheduled in the queue is equal to ν_{j+1}. Based on the no-show rate behavior analyzed in Section 5.2, we

can assume that the no-show probability of the system can be described by a monotonic sequence $\nu_{j-1} \leq \nu_j, j = 1, \ldots, K$. Converted to units of time, the no-show probabilities do reflect the non-increasing temporal trend observed in Section 5.2: $\nu_{\lfloor (j-1)/\mu \rfloor} \leq \nu_{\lfloor j/\mu \rfloor}, j = 1, \ldots, K$. Each patient that does not show up generates a small revenue θ_N, as clinicians can perform other duties that still add value to the system (e.g., clinical notes) [184]. A patient that shows up provides a nominal unit of revenue.

Let $p_j(K)$ be the steady-state probability that upon arrival there are j patients scheduled in the system, and $p_0(K)$ be the steady-state probability that the system is idle, i.e., there are no scheduled patients. Let $\rho = \frac{\lambda}{\alpha}$, $\delta = \frac{\mu}{\alpha}$. As we consider a reversible Markov chain, we can derive the steady-state equations from the local balance equations for the M/M/1/K queuing system [13]:

$$\begin{aligned} p_{j+1}(K) &= \frac{\rho}{\delta + j} p_j(K), \quad j = 0, \ldots, K-1, \\ \sum_{j=0}^{K} p_j(K) &= 1. \end{aligned} \tag{5.1}$$

Then the closed-form expressions of the steady-state probabilities are:

$$p_j(K) = \frac{\rho^j}{\prod_{i=0}^{j-1} (\delta + i)} p_0(K) = \rho^j \frac{\Gamma(\delta)}{\Gamma(j + \delta)} p_0(K), \quad j = 1, \ldots, K, \tag{5.2}$$

with

$$p_0(K) = \frac{1}{1 + \sum_{j=1}^{K} \frac{\rho^j}{\prod_{i=0}^{j-1} (\delta + i)}} = \frac{1}{1 + \Gamma(\delta) \sum_{j=1}^{K} \frac{\rho^j}{\Gamma(\delta + j)}}, \tag{5.3}$$

where $\Gamma(\cdot)$ is the gamma function defined as $\Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t} dt$.

Let $P_S(K)$ be the fraction of patients served, $P_N(K)$ the fraction of no-show patients, $P_C(K)$ the fraction of patients that canceled, and $P_B(K)$ the fraction patients that are blocked. We have the following expressions:

$$\begin{aligned} P_S(K) &= \sum_{j=1}^{K} p_j(K)(1 - \nu_j) \frac{\mu}{\lambda} = \frac{\mu}{\lambda} (1 - p_0(K)) - P_N(K), \\ P_N(K) &= \sum_{j=1}^{K} p_j(K) \beta_j \nu_j \frac{\mu}{\lambda}, \\ P_C(K) &= \sum_{j=1}^{K} p_j(K)(1 - \beta_j) \frac{\mu}{\lambda} = 1 - p_K(K) - \frac{\mu}{\lambda}(1 - p_0(K)), \\ P_B(K) &= p_K(K), \end{aligned}$$

The long-run expected revenue of the system can be formulated as:

$$R(K) = P_S(K) + P_N(K)\theta_N - P_B(K)\theta_B - P_C(K)\theta_C, \tag{5.4}$$

where $0 \leq \theta_N < 1$ since no-shows provide a smaller reward than served patients. The penalties are such that $\theta_B \geq 0$, and $\theta_C \geq 0$. Furthermore, we assume $\theta_N = 0$.

The scheduling interval problem can be formulated as follows:

$$\sup_{K \in \mathbb{Z}^+} R(K). \tag{5.5}$$

From the queueing model, we can derive several structural properties. The analytical details are presented in the Appendix to this chapter. These properties show, among others, that the optimum scheduling interval depends on the parameter settings and patient characteristics. Depending on the settings, several objective function forms can be derived, including forms with quasi-concave revenue functions, as explained in the Appendix.

5.4 Simulation model

To assess the ability of the queuing system to capture reality, we develop a simulation model that captures the time-dependent cancellation behavior of the real system. We need to evaluate the impact of cancellation behavior, as two assumptions were made in the queueing model. A first assumption in the queuing model is that all patients are served FCFS. This implicates that when a cancellation occurs, all patients in line after this canceled appointment will be served one timeslot earlier. However, in practice empty spots due to cancellations are only filled when a new patient arrives that is willing to take that spot. Therefore, some slots might end up empty, if no patient arrives in the interval between the cancellation and the service of this specific appointment slot. A second assumption in the queuing model is that the cancellation rate is exponentially distributed with asymptote 1. However, the data analysis showed that the systems under consideration had lower asymptotes and a bimodal distribution. Therefore, the simulation model captures the time-dependent cancellation behavior of the real system using the empirical distributions, to analyze the impact of these assumptions on the system performance

The discrete event simulation model consists of a single server with a limited buffer of size $K-1$. The buffer represents the available appointment slots, where position 1 equals the slot that is served first, and position $K-1$ the slot that is served last. Together with the server, this makes the total number of positions in the system equal to K.

Patients arrive to the system according to a Poisson distribution with rate λ. Arriving patients are assigned the first available empty position in the buffer. If the buffer is full, patients are rejected.

When the deterministic server becomes empty, it processes the patient at position one in the buffer. If no patient is available at this position (independent

of other possible patients in the queue), the server will remain empty for one timeslot. If a patient is available, and Δt equals the waiting time of this patient in the queue, with probability $\nu_{\Delta t}$ the patient is a no-show, and the server stays empty for one timeslot. With probability $1 - \nu_{\Delta t}$, the patient is seen, and is served. We assume deterministic service times with rate μ, as we consider a tactical level appointment system design.

Patients may cancel their appointment when they are in the buffer. The probability that a patient cancels an appointment depends on the scheduling interval. If a patient cancels, it is timed according to the empirical cancellation timing rates of Section 5.2. The patient departs from the buffer, leaving an empty position in the buffer.

In the simulation model we measure several performance indicators. We record the proportion of rejected, canceled, no-show, and seen patients, as well as the proportion of time the server is idle. This enables a comparison with the queuing system. Furthermore, we register the number of empty slots due to cancellations and an empty system.

We validated the simulation model by comparing the results of this model against the performance in practice. The no-show and cancellation probabilities from the simulation are 10.0 percent and 10.1 percent on average, respectively, which are similar to the actual no-show rate of 10.3percent and cancellation rate of 10.0 percent calculated from historical scheduling data. Therefore, the simulation model is considered valid.

The simulation is developed in Tecnomatix Plant Simulation 11, and simulates 5 years, with a warm-up period of 75 days and 8 replications.

5.5 Experiment design

This section describes the experiments, where we compare the performance of the M/M/1/K queuing system with the simulation performance. First, the base case and experiment settings are described in Section 5.5.1. Second, Section 5.5.2 presents the experiment results.

5.5.1 Base case and experiment settings

We consider a clinic which operates five days a week. Every day, six appointment slots are available. Weekends are excluded from the analysis. As six appointment slots are available per day, we fix the service rate μ to $\mu = 6$. Patients arrive according to a Poisson distribution, with rate $\lambda = 6$. The no-show and cancellation rates are exponentially distributed, and derived from the data-analysis of Section 5.2. We consider the no-show rate of Gallucci et al. [104] (G05) as the base case no-show rate, as G05 has been used in the literature most frequently [116, 184]. Furthermore, patients cancel their appointments with rate $\alpha = 0.06$, as derived from Institution 1, corresponding with the cancellation rate in Figure 5.1.

As all cost parameters are normalized towards the revenue from serving one patient in one timeslot, we need to assess the cost of cancellation (θ_C) and the cost of rejection (θ_B). As cancellations have a higher impact on the system (i.e., through an extra administrative burden, blocking slots for patients that would have showed up), we expect the cost of cancellation to be higher than the cost of rejection.

As we expect rejected patients to be booked in another clinic, or be overbooked in non-clinic hours, which is the current practice in both hospitals included in this research, we do not consider a cost of lost patients for rejected patients, but we do include an inconvenience cost. Canceled patients however might end up being lost by the clinic, as not every patient will reschedule their appointment.

As there is a tradeoff between rejection and cancellation, decision makers should together decide upon the cancellation and rejection cost parameters, based on the aforementioned considerations. Therefore, we experiment with various cost settings as shown in Table 5.4, to analyze the tradeoff between cancellations, rejections, and serving patients. In the base case we use the settings θ_B=1.2 and θ_C=1.4.

To evaluate the efficiency of the method and to assess the behavior of various system settings, we execute the following experiments, as shown in Table 5.4. First, we analyze the impact of the no-show and cancellation rate on the optimal scheduling interval. Seven no-show rates are considered, five derived from the literature and two derived from hospital data (refer to Section 5.2). Although many studies report on the time dependency of no-show rates, most literature does not include a functional form of the time dependent no-show rate which is based on real-life data [116]. Furthermore, most literature does not force their rates to long term asymptotic behavior, despite the fact that both no-show and cancellation probabilities are not allowed to exceed one. Therefore, we limit our literature rate inclusion to rates that are monotonically increasing and converging towards a maximum value, which does not exceed one. We were able to identify five studies that provided such measurements over multiple scheduling intervals, for which we can derive a functional form. Their parameters are presented in Table 5.3.

Table 5.3 Parameter settings for literature based no-show rates per scheduling interval in days

Study	Name	$\nu_{\max}$	ν_0	C
Benjamin-Bauman et al. 1984	BB84	0.48	0.16	7
Festinger et al. 2002	F02	0.67	0.05	2
Gallucci et al. 2005	G05	0.43	0.11	2
Green and Savin 2008	GS08	0.31	0.01	50
Whittle et al. 2008	W08	0.21	0.11	6

For the cancellation rate we explore the system's behavior with three rates,

varying around the rate derived from the data. As functional forms of the cancellation rate are rarely reported upon in the literature, we identified only one manuscript provides cancellation measures for multiple scheduling intervals [321]. They found a similar monotonic relationship for patient initiated as well as clinic initiated cancellations. The exponential function parameters for the cancellation rate of Whittle et al. [321] are $\chi_{\max} = 0.24$, $\chi_0 = 0.09$, and $C = 10$, which can be approximated with $\alpha = 0.05$, and is included in the experiments as well.

Since some studies include cancellations in the no-show rates, we should be careful with the comparison of the various rates derived from these studies with our data-driven rates. However, they are valuable for analysis, since late cancellations may end up as empty appointment slots, and therefore reflecting no-show behavior.

No study reported cancellation timing measures. Therefore, we base the timing behavior on the observations in the data analysis. In the analytical model we use an exponential distribution to determine the cancellation timing, whereas in the simulation, the cancellation rates from Institutions 1 and 2 have an empirically distributed timing distribution dependent on the scheduling interval based on the observations in Section 5.2.

Besides analyzing the impact of the no-show and cancellation rate, we also analyze the impact of the arrival rate on the clinic behavior. In line with Liu [184], we expect higher arrival rates to result in lower scheduling windows, and vice versa. Third, we consider multiple combinations of the cost coefficients θ_C and θ_B, to analyze the effect for various system settings. Fourth, we perform two case study experiments, with data from Institutions 1 and 2, to analyze the performance of our model on real life data, and to assess if the model is generalizable in practice. The case study of Institution 2 uses the corresponding no-show rate from Table 5.2, and a cancellation rate of $\alpha = 0.032$, as derived from Figure 5.2.

Considering the aforementioned parameters, we obtain a base case and 19 experiment instances. Table 5.4 gives an overview of the instances.

5.5.2 Experiment results

Table 5.5 provides an overview of the results of the queuing model experiments. The first ten experiments show the impact of the no-show and cancellation rates. Here, it is shown that for various no-show rates, an infinite queue is optimal. These no-show rates have amongst the lowest asymptotes considered in the experiments, which supports the hypothesis that the lower the impact of no-shows, the longer the queue can be. The impact of the cancellation rate to the optimal scheduling interval is less clear. A small increase in queue length can be observed for lower cancellation rates, but no statistically significant difference is observed between the performance of the subsequent experiments. In additional experiments (not reported), we observe that low-traffic systems are more sensitive to no-show and cancellation behavior of patients.

Experiments 11 to 14 evaluate the impact of the arrival rate. Table 5.5 shows

Table 5.4 Input parameter variations for the experiments

Exp no.	μ	λ	No-show rate	Canc. rate	(θ_B, θ_C)
Base case	6	6	G05	0.06	(1.2, 1.4)
1	6	6	BB84	0.06	(1.2, 1.4)
2	6	6	F02	0.06	(1.2, 1.4)
3	6	6	GS08	0.06	(1.2, 1.4)
4	6	6	W08	0.06	(1.2, 1.4)
5	6	6	Inst. 1	0.06	(1.2, 1.4)
6	6	6	Inst. 2	0.06	(1.2, 1.4)
7	6	6	G05	0.10	(1.2, 1.4)
8	6	6	G05	0.075	(1.2, 1.4)
9	6	6	G05	0.05	(1.2, 1.4)
10	6	6	G05	0.025	(1.2, 1.4)
11	6	5	G05	0.06	(1.2, 1.4)
12	6	7	G05	0.06	(1.2, 1.4)
13	6	8	G05	0.06	(1.2, 1.4)
14	6	10	G05	0.06	(1.2, 1.4)
15	6	6	G05	0.06	(1.1, 1.5)
16	6	6	G05	0.06	(0.8, 0.9)
17	6	6	G05	0.06	(0.8, 1.2)
18	6	6	G05	0.06	(1, 1)
19	6	6	Inst. 2	0.032	(1.2, 1.4)

a decrease in optimal scheduling window for higher values of λ. Thus, for high demand systems, it is beneficial to reduce the scheduling window, and possibly organizing the clinic on a walk-in basis. This ensures that as many patients as possible can be served, as the patients that make an appointment, will most likely not end up as a no-show or cancellation. This corresponds to the finding of Liu [184].

Experiments 15 to 18 evaluate the impact of the cost coefficients on the scheduling window. We observe that when provider idle time is more important to the decision makers than rejections, the scheduling interval is shorter than when idle time and rejections are equally valued. Therefore, the optimal scheduling window depends on the weights which decision makers assign to the cost coefficients, such as rejecting patients or provider idle time.

The case study experiments show that both for Institution 1 (experiment 5) and Institution 2 (experiment 19) an infinite scheduling window is optimal.

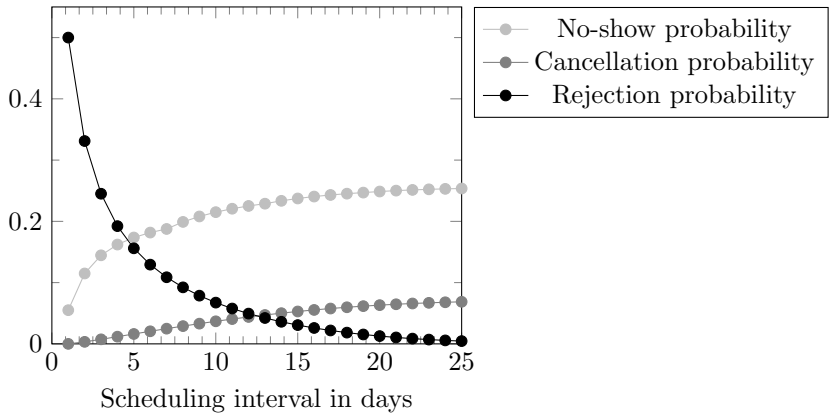
In all experiments, the optimal scheduling window is found through a trade-off between no-shows and cancellations, and patient rejections. For the base case, this is visualized in Figure 5.5. As expected, this figure shows that the no-show and cancellation probabilities increase with longer scheduling windows, as patients are allowed to have longer waiting times. The rejection probability decreases with longer scheduling windows, as more patients are admitted in the system.

Table 5.5 Experiment results

Exp no.	K^*	Days	Obj. value
Base case	21	3	3.430
1	∞	∞	3.701
2	13	2	2.995
3	∞	∞	4.851
4	∞	∞	4.218
5	∞	∞	4.464
6	∞	∞	4.420
7	20	3	3.288
8	21	3	3.374
9	22	3	3.468
10	23	3	3.568
11	∞	∞	3.507
12	13	2	2.692
13	7	1	1.693
14	7	1	-0.408
15	19	3	3.401
16	19	3	3.651
17	13	2	3.549
18	25	4	3.599
19	∞	∞	4.666

A simulation study is done to evaluate the effects of neglecting the timing of cancellations on the results. For each of the experiments, we simulated the system with the corresponding K^* from Table 5.5. In the simulation the average percentage of idle time over all experiments was 26.2% (21.3% due to no-shows,

Figure 5.5 Average no-show, cancellation and rejection probabilities per scheduling interval



and 4.9% due to an empty system). In the analytical results, the average idle time over all experiments was 26.9% (19.5% due to no-shows, and 7.4% due to an empty system).

The simulation shows that the number of empty slots in the queueing model is overestimated, as the system is 0.7% of the total time less idle. In the queueing model the idle time due to no-shows is underestimated in all analytical experiments, although the percentage of time the system is empty is overestimated on average. Only few simulation experiments showed higher overall idle system probabilities, as expected due to the cancellation timing. For example simulation experiments 12 to 14 showed higher system emptiness compared to the analytical results. In these experiments, the system was overloaded with patients, which makes an empty system highly unlikely in the analytical model given the FCFS assumption. Therefore, the increase is primarily due to the impact of late cancellations. The highest idle times in both experiment settings are seen in experiment 11, as there are often no patients in the system, since the average number of arrivals is lower than the capacity. The lowest idle times are seen in experiment 3, due to its low no-show rate.

5.6 Conclusions and discussion

No-show and cancellation behavior of patients influence the performance of hospital's outpatient clinics. As less than 50% of all scheduled appointments result in an actual patient being seen by the specialist, clinics face a significant problem. We investigated the scheduling interval in relation to no-show and cancellation rates, and found that an increasing scheduling interval results in higher no-show and cancellation probabilities. Therefore, clinics can benefit from limiting the scheduling interval using a scheduling window, to minimize the negative effect of no-shows and cancellations. The optimal scheduling window is found through a tradeoff between the price of cancellations and no-shows and the price of rejection.

We developed an analytical queuing model to determine this optimal scheduling window, and provided a simulation study to evaluate the effectiveness of this model. Our results show that for systems with a relatively high number of arrivals, as shown in experiments 12-14, it is beneficial to limit the scheduling window. The impact of the no-show and cancellation rate showed to have a large impact on the optimal scheduling window in low-traffic systems. A limited scheduling window is also preferred for systems that highly value the utilization of the providers, as shown in experiment 17. Note that for systems with an infinite scheduling window, it is still beneficial to schedule patients as early as possible, as this maximizes the probability that the patient will show up for the appointment.

Systems with low no-show and cancellation rates should increase their scheduling window in order to prevent unnecessary rejections. The simulation study showed that the assumption of the timing of the cancellations to be exponen-

tially distributed, gives a good approximation of the expected results.

5.6.1 Contributions

Our research provides multiple contributions. First, we show that the no-show and cancellation rates are time-dependent. A longer scheduling window results in higher no-show and cancellation probabilities.

Second, not only the occurrence of cancellations is related to the scheduling interval, but also the timing of cancellations. We show that cancellation timing follows a bimodal distribution, where peaks in cancellations are observed right after the creation of the appointment, and just before the actual appointment date. This corresponds with the literature that analyzed reasons for cancellations, where scheduling conflicts, forgetting the appointment, and logistical challenges are frequently observed as main reasons for patient cancellations.

Third, we develop an analytical model to incorporate the time-dependent no-show and cancellation rates in the design of an appointment system.

Fourth, through an extensive simulation analysis we show that this analytic model is a good representation of reality.

Fifth, we showed the general applicability of this model by case studies of outpatient clinics of two hospitals in different health systems. Although in these two case studies an infinite scheduling window was determined to be optimal, for certain combinations of no-show and cancellation rates derived from real-world scenarios, significant efficiency gains can be achieved, when a limited optimal scheduling window is used. This shows that limiting the scheduling window can be a means to mitigate the effect of no-shows and cancellations.

Sixth, our data-analysis and model provide insight into the impact of no-shows and cancellations. Where clinics tend to put more emphasis on reducing the number of no-shows compared to cancellations, this researched showed that when focusing on the scheduling interval, the number of cancellations should get more attention, as the scheduling interval dependent no-show rate converges faster than the cancellation rate. Therefore, more efficiency gains can be derived in reducing the number of cancellations.

Seventh, we show that for low demand and low no-show and cancellation clinics, it is optimal to have a long scheduling window, whereas for high demand and high no-show and cancellation clinics the optimal scheduling window is as short as possible.

5.6.2 Further research

From this research, we observe multiple areas for further research.

It is unknown how reminders and penalties for no-show impact the bimodal distribution of cancellation timing, and how this impacts the optimal scheduling window. As we hypothesize that more patients will cancel their appointment right before the actual appointment, the possibilities of reallocating slots to new arrivals will decrease, and more canceled slots will end up idle.

Literature has shown that new patients are more sensitive for long scheduling intervals than established patients [81]. Possible reasons are (un)established relationships with their clinicians, and shopping around possibilities. As very large datasets are required to define reliable time-dependent no-show and cancellation behavior for subgroups, such as new and established patients, further research in large healthcare institutes with reliable data collection systems, is required to enable subgroup analyses.

Further research in the implementation of short scheduling windows is also required. Patients may want an appointment further in the future than the optimal scheduling window allows for, or cannot be scheduled due to the short scheduling window. One possibility is to organize the clinic on a walk-in basis. An alternative is to maintain a call list. In such a system, a patient, who was not given an appointment within the scheduling window, would be added to this list and called to arrange an appointment one scheduling window is extended. Another alternative is to implement a carefully designed admission control policy to reject patients. Our hospitals provide patients, who would normally be rejected due to completely booked calendars, an appointment slot in overtime. Another policy could be to refer the patient to a partnering clinic. Each of these interventions can ensure that as many patients as possible are served, as an appointment scheduled within a shorter scheduling window is less likely to result in a no-show or cancellation.

Also, the implementation of the rejection policy should be thought through. Our hospitals provide patients who will be rejected an appointment slot in overtime, which is a possible way to deal with the patients that were otherwise rejected. Another policy could be to refer the patient to a partnering clinic.

To enable computational efficiency, our model is stylized. For implementation in practice, further research is needed to analyze the effect of patient choice and the exponential cancellation behavior. We assumed an FCFS policy, but in practice, patient preferences are highly diverse and complex, as patients do not necessarily prefer the first available appointment slot. However, none of the collaborating hospitals have reliable data to determine slot preference probability functions. However, operations managers at both hospitals feel they are able to accommodate most requests while maintaining a high level of planned slot utilization. Hence, their realized schedule roughly resembles the result of an FCFS queueing discipline. An important future research direction is to derive reliable slot preference functions to improve the validity of the model.

Furthermore, we fit an exponential distribution to the cancellation data. However, this does not incorporate the bimodal behavior of the cancellation timing, as it assumes that the cancellation probability is higher when a longer waiting list is observed. Our simulation study showed that the real-life cancellation behavior did not influence the outcomes of our analytic model. However, further research is required to assess the effects of other operational system behavior in a similar fashion.

Further research is required in the cancellation behavior of patients and institutions, to further distinguish the various cancellation types. Also, immediate

cancellations and late cancellations should be studied, ideally to be able to include these two cancellation types as individual rates to increase the validity of the model.

No-show and cancellation behavior not only influence the scheduling window. Further research in incorporating these rates and the bimodal cancellation timing distribution in the design of (other elements of) appointment systems is required.

5.7 Appendix I

This appendix provides the analytical results of the structural properties presented in Section 5.3. In this section the notation $q_K(\cdot)$ is used to represent, as a function of queue length, the steady-state probability of rejection, i.e., the length of the queue is at full capacity. Let us notice that the functions $p_0(K)$ and $q_K(K)$ are defined in $\mathbb{Z}^+$, whereas the function $p_j(K)$ is defined in $\mathbb{Z}_j^+ := \{K \in \mathbb{Z}^+ | K \geq j\}$.

The monotonic properties of the steady-state probabilities with respect to length of the queue K are summarized below.

Lemma 1. *Given $K_1, K_2 \in \mathbb{Z}^+$ such that $K_1 \leq K_2$ then $q_K(K_1) \geq q_K(K_2)$, and $p_j(K_1) \geq p_j(K_2)$ for $j \in \mathbb{Z}^+$.*

Proof. Let $K_1, K_2 \in \mathbb{Z}^+$ such that $K_1 \leq K_2$, then:

$$\sum_{j=0}^{K_1} \Gamma(\delta) \frac{\rho^j}{\Gamma(\delta + j)} \leq \sum_{j=0}^{K_2} \Gamma(\delta) \frac{\rho^j}{\Gamma(\delta + j)},$$

since the summation involves non-negative numbers. It follows from (5.3) that $p_0(K_1) \geq p_0(K_2)$.

The recursive steady-state equations (5.1) shows that, for $j \in \mathbb{Z}^+$, the function $p_j(K)$ is non-increasing in its respective domain. Finally, we will show that the probability of rejection is non-increasing. Let $K \in \mathbb{Z}^+$, then using the closed-forms (5.2)-(5.3) we have:

$$\begin{aligned}
& q_K(K) - q_K(K+1) \\
= \; & \frac{\rho^K / \Pi_{i=0}^{K-1}(\delta + i)}{1 + \sum_{j=1}^{K} \rho^j / \Pi_{i=0}^{j-1}(\delta + i)} - \frac{\rho^{K+1} / \Pi_{i=0}^{K}(\delta + i)}{1 + \sum_{j=1}^{K+1} \rho^j / \Pi_{i=0}^{j-1}(\delta + i)} \\
= \; & \frac{\rho^K}{\Pi_{i=0}^{K-1}(\delta + i)} \left(\frac{1}{1 + \sum_{j=1}^{K} \rho^j / \Pi_{i=0}^{j-1}(\delta + i)} - \frac{\rho / (\delta + K)}{1 + \sum_{j=1}^{K+1} \rho^j / \Pi_{i=0}^{j-1}(\delta + i)} \right) \\
= \; & \frac{\rho^K}{\Pi_{i=0}^{K-1}(\delta + i)} p_0(K) p_0(K+1) \left(1 + \sum_{j=1}^{K-1} \rho^j \left(\frac{1}{\delta + j} - \frac{1}{\delta + K} \right) \right) \geq 0.
\end{aligned}$$

□

Let $\gamma(x, a)$ be the normalized lower incomplete gamma function defined as:

$$\gamma(x, a) = \frac{1}{\Gamma(a)} \int_0^x e^{-t} t^{a-1} dt.$$

Using the function $\gamma(\cdot, \cdot)$ we can reformulate the closed-form (5.3) as Ancker and Gafarian [13]:

$$p_0(K) = \left[1 + e^{\rho} \rho^{1-\delta} \Gamma(\delta)(\gamma(\rho, \delta) - \gamma(\rho, \delta + K))\right]^{-1}.$$

It follows from the expression above and Lemma 1 that:

$$\lim_{K \to +\infty} p_0(K) = P_0 = \left[1 + e^{\rho} \rho^{1-\delta} \Gamma(\delta) \gamma(\rho, \delta)\right]^{-1}, \tag{5.6}$$

$$\lim_{K \to +\infty} q_K(K) = 0, \tag{5.7}$$

$$\lim_{K \to +\infty} p_j(K) = \frac{\rho^j \Gamma(\delta)}{\Gamma(\delta + j)} P_0, \quad j \in \mathbb{Z}^+. \tag{5.8}$$

The limits shown in (5.7) - (5.8) result from that the gamma function grows faster than any power function.

Lemma 2. *The rejection probability* $\{q_K(K)\}_{K \in \mathbb{Z}^+}$ *and* $\{p_j(K)\}_{K \in \mathbb{Z}_j^+}$*, for* $j \in \mathbb{Z}^+ \cup \{0\}$*, are convex sequences.*

Proof. A sequence is convex if its first difference is non-decreasing. Let $j = 0$, define the first difference sequence $\{m_k\}_{k \in \mathbb{Z}^+}$ as:

$$m_k = p_0(K+1) - p_0(K) = -\rho^{K+1} \frac{\Gamma(\delta)}{\Gamma(\delta + K + 1)} p_0(K) p_0(K+1).$$

We need to show that $m_k \leq m_{k+1}$, i.e, $m_k - m_{k+1} \leq 0$:

$$m_k - m_{k+1} = \rho^{K+1} \frac{\Gamma(\delta)}{\Gamma(\delta + K + 1)} p_0(K+1) \left(\frac{\rho}{\delta + K + 1} p_0(K+2) - p_0(K) \right). \tag{5.9}$$

By Lemma 1 we know that $q_K(\cdot)$ is non-increasing then:

$$q_K(K+1) - q_K(K) = \rho^K \frac{\Gamma(\delta)}{\Gamma(\delta + K)} \left(\frac{\rho}{\delta + K} p_0(K+1) - p_0(K) \right) \leq 0,$$

therefore:

$$\frac{\rho}{\delta + K} p_0(K+1) - p_0(K) \leq 0. \tag{5.10}$$

Using (5.10) in (5.9) we have:

$$m_k - m_{k+1} \leq \rho^{K+1} \frac{\Gamma(\delta)}{\Gamma(\delta + K + 1)} p_0(K+1) \left(p_0(K+1) - p_0(K) \right) \leq 0,$$

hence $\{p_0(K)\}_{K\in\mathbb{Z}^+}$ is a convex sequence. The convexity of $\{p_j(K)\}_{K\in\mathbb{Z}_j^+}$ follows from (5.1).

Finally, let η_k be the first difference of the rejection probability sequence:

$$\begin{aligned}\eta_k &= q_K(K+1) - q_K(K) \\ &= -\frac{\rho^K}{\Pi_{i=0}^{K-1}(\delta+i)} p_0(K)p_0(K+1)\left(1+\sum_{j=1}^{K-1}\rho^j\left(\frac{K-j}{(\delta+j)(\delta+K)}\right)\right),\end{aligned} \tag{5.11}$$

then:

$$\begin{aligned}\eta_k - \eta_{k+1} = q_K(K+1)p_0(K+2)&\left(1+\sum_{j=1}^{K}\frac{\rho^j(K-j+1)}{(\delta+j)(\delta+K+1)}\right) \\ -\frac{\delta+K}{\rho}q_K(K+1)p_0(K)&\left(1+\sum_{j=1}^{K-1}\frac{\rho^j(K-j)}{(\delta+j)(\delta+K)}\right),\end{aligned} \tag{5.12}$$

using the result (5.10) and the closed-forms (5.2) in equation (5.12) we get, after algebraic manipulations, that $\eta_k - \eta_{k+1} \leq 0$, which shows the convexity of $\{q_K(K)\}_{K\in\mathbb{Z}^+}$. □

Expanding the term in (5.4), the revenue function $R(K)$ can be expressed as $R(K) = \lambda T(K) - \lambda\theta_C$ with:

$$T(K) = \frac{\mu}{\lambda}(1-p_0(K))(1+\theta_C) + P_N(K)(\frac{\mu}{\lambda}\theta_N - 1) + q_K(K)(\theta_C - \theta_B). \tag{5.13}$$

As stated in the previous section, the no-show probabilities of the system are described by a sequence $\{\nu_j\}_{j\in\mathbb{Z}^+}$ such that $\nu_j \leq \nu_{j+1}$ for all $j \in \mathbb{Z}^+$ and $\lim_{j\to+\infty}\nu_j = \nu^*$, $\nu^* \in [0,1]$. Then, $P_N(K)$ is bounded for all $K \in \mathbb{Z}^+$ and $\lim_{K\to+\infty} P_N(K) \leq \frac{\mu}{\lambda}(1-P_0)\nu^*$, because from (5.2) we have:

$$P_N(K) = \sum_{j=0}^{K-1} p_j(K)\beta_j\nu_j = \frac{\mu}{\lambda}\sum_{j=1}^{K} p_j(K)\nu_{j-1} \leq \frac{\mu}{\lambda}(1-p_0(K))\nu_{K-1}, K \in \mathbb{Z}^+. \tag{5.14}$$

In order to gain some insights of the structure of the problem, we will consider the particular case $\nu_j = \nu, j \in \mathbb{Z}^+$. In this case, the function $T(K)$ has a simple form:

$$T(K) = T_\nu(K) = \frac{\mu}{\lambda}(1-p_0(K))(1+\theta_C+(\frac{\mu}{\lambda}\theta_N - 1)\nu) + q_K(K)(\theta_C - \theta_B). \tag{5.15}$$

Lemma 3. *If $\theta_B \geq \theta_C \geq 0$ then $T_\nu(K)$ is increasing in the domain $\mathbb{Z}^+$.*

Proof. Let $K_1, K_2 \in \mathbb{Z}^+$ such that $K_1 \leq K_2$, then:

$$\begin{aligned} T_\nu(K_1) - T_\nu(K_2) &= \frac{\mu}{\lambda}(1+\theta_C+(\frac{\mu}{\lambda}\theta_N-1)\nu)(p_0(K_2)-p_0(K_1)) \\ &\quad +(\theta_C-\theta_B)(q_K(K_1)-q_K(K_2)), \\ T_\nu(K_1) - T_\nu(K_2) &\leq (\theta_C-\theta_B)(q_K(K_1)-q_K(K_2)), \\ T_\nu(K_1) - T_\nu(K_2) &\leq 0, \end{aligned}$$

where the term $(1+\theta_C+(\frac{\mu}{\lambda}\theta_N-1)\nu) \geq 0$ since $0 \leq \theta_N < 1$. The first inequality follows from the decreasing property of $p_0(K)$. The last inequality is obtained from Lemma 1 and the condition $\theta_B \geq \theta_C$. □

An implication of Lemma 3 is that the function $T_\nu(K)$ does not have a maximum in $\mathbb{Z}^+$ since

$$\sup_{K\in\mathbb{Z}^+} T_\nu(K) = \lim_{k\to+\infty} T_\nu(K) = \frac{\mu}{\lambda}(1-P_0)(1+\theta_C+(\frac{\mu}{\lambda}\theta_N-1)\nu). \tag{5.16}$$

Therefore, the scheduling interval of the system can be as large as possible if the probabilities of no-shows behave relatively constant with respect to the capacity of the queue, and there is a preference to set up a higher penalty for blocking patients.

Another insight of Lemma 3 is that if the function $T_\nu(K)$ has a maximum in $\mathbb{Z}^+$, then:

$$\max_{K\in\mathbb{Z}^+} T_\nu(K) > T_\nu^*,$$

where $T_\nu^* = \frac{\mu}{\lambda}(1-P_0)(1+\theta_C+(\frac{\mu}{\lambda}\theta_N-1)\nu)$.

Consequently, if $0 < \theta_B < \theta_C$ we can truncate the domain of $T_\nu(K)$ by selecting a small tolerance number $\tau > 0$ to find the smallest $\bar{K} \in \mathbb{Z}^+$ such that $p_0(K) - P_0 < \epsilon$ and $q_K(K) < \epsilon$ for all $K \geq \bar{K}$, where $\epsilon = \tau/2((1+\theta_C+(\frac{\mu}{\lambda}\theta_N-1)\nu)+\theta_C-\theta_B)$. Then, the following optimization problem always has a solution, and it can be solved by enumeration:

$$\max_{K\in\{1,\ldots,\bar{K}\}} T_\nu(K). \tag{5.17}$$

Let us notice that, if $\bar{K}$ is a solution of (5.17) then $T_\nu(K)$ does not have a maximum in $\mathbb{Z}^+$, because:

$$\begin{aligned} |T_\nu(\bar{K}) - T_\nu(K)| &\leq |T_\nu(\bar{K}) - T_\nu^*| + |T_\nu(K) - T_\nu^*|, \\ &< \tau, \text{ for all } K \geq \bar{K}. \end{aligned}$$

In addition, by Lemma 2, problem (5.17) is a difference of convex (DC) optimization problem. Therefore, the existence of a solution of (5.17) such that $K < \bar{K}$, depends on the decrease rate of the functions $p_0(\cdot)$ and $q_K(\cdot)$. For example, if $\delta \leq \rho$ and $\alpha < \mu$ then we can find $\tilde{K} \in \mathbb{Z}^+$ such that:

$$\frac{\rho^K\Gamma(\delta)}{\Gamma(\delta+K)} \leq \frac{\rho^{\tilde{K}}\Gamma(\delta)}{\Gamma(\delta+\tilde{K})}, \text{ for all } K \leq \tilde{K},$$

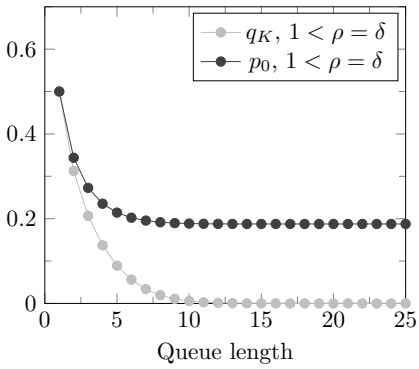


Figure 5.6 Steady-state probability of rejection and idle for $\mu = \lambda = 6$, $\alpha = 0.6$

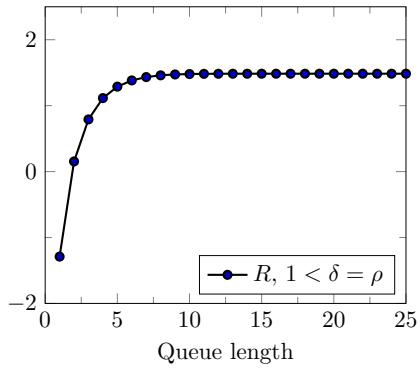


Figure 5.7 Revenue function with $\theta_N = 0$, $\nu = 0.43$, $\theta_B = 1.0$, $\theta_C = 1.15$

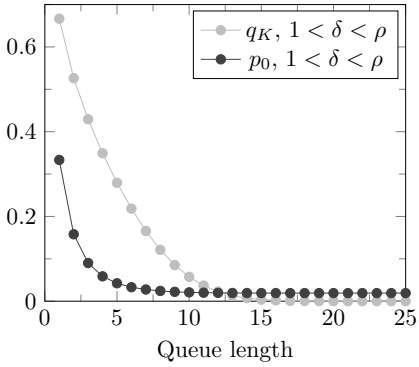


Figure 5.8 Steady-state probability of rejection and idle for $\mu = 3$, $\lambda = 6$, $\alpha = 0.6$

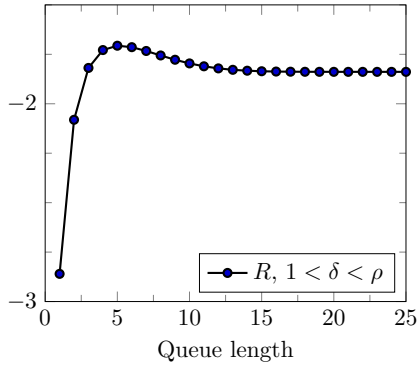


Figure 5.9 Revenue function with $\theta_N = 0$, $\nu = 0.43$, $\theta_B = 1.0$, $\theta_C = 1.15$

since $\rho > 1$ and the function $\Gamma(\delta)/\Gamma(\delta+K)$ is decreasing in $\mathbb{Z}^+$ if $\delta > 1$. Therefore, $q_K(K) \geq p_0(K)$ for all $K \leq \tilde{K}$, and $q_K(K) \leq p_0(K)$ for all $K > \tilde{K}$, but the difference between these values is not necessarily monotonic. Figure 5.6 shows that the difference of q_K and p_0 is non-decreasing, which could produce a non-decreasing revenue function, as displayed in Figure 5.7. If δ is reduced so that P_0 is a small number, the difference of q_K and p_0 has an interesting behavior as shown in Figure 5.8. It can be observed that the difference starts to decrease for values of $K \geq 5$, approximately. This behavior could define a quasi-concave revenue as illustrated in Figure 5.9.

Finally, for a general form of $P_N(K)$ we still can solve the problem by enumeration as in (5.17), because $P_N(K)$ has a horizontal asymptote. In addition, by (5.14) the function $T(K)$ is dominated by a function that behaves like $T_\nu(K)$.

CHAPTER 6

Stochastic integer programming for multi-disciplinary outpatient clinic planning

6.1 Introduction

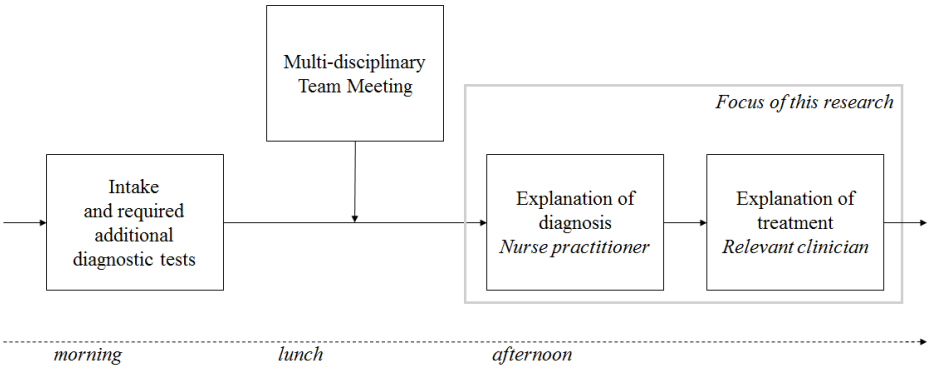
During the redesign of one of UMC Utrecht's cancer outpatient clinics, a decision on the blueprint of the agendas of the involved nurse practitioners and clinicians has to be made. This is a complex decision, as multiple patient types are involved, and the overall performance of the cancer clinic depends on the interplay between all agendas. Therefore, the optimization of the blueprint schedules of this multi-disciplinary clinic requires an integrated optimization approach, in which all appointment schedules are jointly optimized.

As seen in Chapter 2, multi-disciplinary teams are increasingly introduced in various medical contexts, such as in outpatient clinics and operating rooms [182, 218], and in various medical disciplines, such as cancer care, rehabilitation, and neurology [113, 262, 301, 302]. However, the coordination and control of these teams is complex, as multiple clinicians from multiple departments are involved.

The contribution of this chapter is that we design optimized blueprint schedules for multi-disciplinary appointment planning at a tactical level of control, which incorporates uncertainties in patient routing. As this currently is an open question in the literature, our research is the first to address this problem. Also, we test the suitability of the approach for the hospital's problem at hand, we compare our results with the current hospital schedules, and present the associated savings. Furthermore, although initiated from a specific cancer clinic application, many other multi-disciplinary applications can benefit from a solid approach towards multi-disciplinary clinic blueprint planning.

This chapter is organized as follows: First, we introduce the problem in Section 6.2. Then, the relevant literature on open access multi-appointment planning is described in Section 6.3. Section 6.4 presents the mathematical problem description and solution method. Section 6.5 presents the proposed solution methodology, followed by the experiments and a case study in Section 6.6 and Section 6.7, respectively. Finally, Section 6.8 gives the conclusions, discussion

Figure 6.1 Diagnostic pathway of a multi-disciplinary cancer patient



and opportunities for further research.

6.2 Problem description

Figure 6.1 shows the pathway of a cancer patient following the diagnostic trajectory in the hospital at hand on an arbitrary day in which a multi-disciplinary team meeting (MTM) takes place. In our collaborating hospital this is every Tuesday. These patients, with (a high probability of having) cancer, are often referred from other hospitals, and require multiple disciplines to be involved in their treatment. Therefore, this pathway starts with an intake, and if required some additional diagnostic tests, followed by an MTM. After the MTM, on that same day, the patient gets two consultations in the multi-disciplinary clinic. The first consultation is with a nurse practitioner (NP) (or another clinician, depending on the preference of the care system), where the patient receives the cancer diagnosis. Thereafter, the patient has a second consultation with a clinician who explains more about the proposed treatment. Each possible treatment is executed by a discipline, with corresponding clinicians who provide the treatment consultation. The treatment modality, and thus the type of clinician needed, is only known during the MTM. Therefore, there is uncertainty about the number of patients that require a consultation for each clinician type. In this chapter we focus on these two consultations, which we will refer to as the *'multi-disciplinary clinic'*. The hospital aims to minimize waiting time between these two consultations, as patients receive a high-impact message from their care providers. Furthermore, the hospital wants their clinicians to be fully utilized. Therefore, the clinicians' overtime and idle time need to be minimized as well.

The patients that follow this care pathway are referred to as *'multi-disciplinary patients'*. Thus, multi-disciplinary patients are patients that require an appointment with an NP, followed by a walk-in appointment with a clinician on a First Come First Serve (FCFS) basis. These multi-disciplinary patients are diagnosed for a specific tumor type. Similar to practice, the schedule for the

NPs is made directly following the MTM, thus the number of multi-disciplinary patients and the treatment modality for every patient is known at the time of scheduling their appointments with the NP. Therefore, the referral probabilities for a multi-disciplinary patient from the NP to the various clinician types are known. Furthermore, since multi-disciplinary patients are already in the hospital, the no-show rate of multi-disciplinary patients is close to 0%. Therefore, we assume all multi-disciplinary patients will show up for their appointment with the NP.

Next to the multi-disciplinary patients, another patient type is admitted in the multi-disciplinary clinic, which we refer to as *'regular patients'*. Regular patients only require a pre-booked appointment with a specific clinician type, for example a check-up appointment. These appointments are booked several weeks to months in advance. The regular patient demand is assumed to be sufficient to fill the maximum capacity of the clinic. We assume all regular patients to show up and to arrive on time for their appointment. Furthermore, they will be served on the time of their appointment, even if a multidisciplinary patient is waiting longer, as pre-booked appointments are prioritized. Since regular patients book their appointment in advance, the regular patient demand is known well before the multi-disciplinary patient demand. Therefore, schedulers need to know to which appointment slots in the clinicians' agendas they can schedule regular patients, as selecting the wrong slots might lead to unnecessary idle, waiting, and overtime.

We aim to derive a planning method for scheduling multi-disciplinary patients in the agenda of the NPs, and a blueprint schedule for each of the clinicians which differentiates between slots for multi-disciplinary patients and regular patients. The NP schedules result in walk-in rates to the various clinician types. As various combinations of patients might result in various NP schedules, the clinicians' blueprint schedules should be optimized together with all possible NP schedules.

For the agenda of the NPs we assume that the number of NPs, and thus the number of available appointment slots per time slot is known, that overbooking is not allowed, and that all slots are booked during every planning period. For the agenda of the clinicians, we assume that the number of clinicians per type are known, and that multi-disciplinary patients that walk-in into the waiting room of a clinician type, wait until the first available empty slot with any of the clinicians of that specific type. Regular patients are always served at the time of their appointment, and double-booked appointment slots are not allowed. We assume that all patients are served, if needed in overtime, as one clinician per clinician type can work in overtime. Furthermore, as all clinicians agreed on the same service duration for all patients, no differentiation between service times of clinician types is required. The blueprint appointment schedule is designed as the number of appointments in the agenda of a clinician that can be booked for a regular patient for each time slot.

To evaluate the performance of the blueprint schedules, multiple objectives should be considered [196]. We consider the optimal schedule to be a schedule that minimizes a cost function, considering the expected multi-disciplinary pa-

tient waiting time between the two multi-disciplinary appointments, the clinician overtime, and the clinician idle time, similar to the cost function considered in [233]. The cost function is influenced by the number of regular patients to be scheduled in the clinicians' schedules and their timing, as well as by the sequence in which multi-disciplinary patients are seen by the NPs.

As an example, consider a very small multi-disciplinary clinic. Here, one NP has 4 time slots, and two clinicians of two clinician types (a surgeon and a medical oncologist) both have 5 time slots, as shown in Figure 6.2. In this clinic, there are multi-disciplinary patients consulted with two tumor types. Multi-disciplinary patients with tumor *dark-gray* account for $1/4^{\text{th}}$ of all multi-disciplinary patients seen, and have a probability of getting surgery of 20%, and a probability of getting chemotherapy of 80%. Multi-disciplinary patients with tumor *white* account for $3/4^{\text{th}}$ of the multi-disciplinary patient population, and have a probability of getting surgery of 60%, and a probability of getting chemotherapy of 40%. The question is how many and in which time slots the clinicians can see regular patients, in order to minimize the expected waiting for multi-disciplinary patients, and to minimize the idle and overtime. Since regular patients want to get their appointment dates multiple weeks in advance, this schedule should be designed before the treatment modalities of the multi-disciplinary patients are known, as their treatment is decided during the MTM. However, at this point we do not know the number of arrivals of the two multidisciplinary patient types. Therefore, all possible optimal schedules of the NP should be taken into account as well, since these schedules determine the arrival rate to the clinicians. Following the MTM, after the treatment modalities of the multi-disciplinary patients are known, the optimal schedule for the NP can be determined and immediately executed, whereas the schedule for the clinicians is already fixed at that moment. An example of a possible schedule for the NP, and a possible blueprint schedule for the clinicians is shown in Figure 6.2.

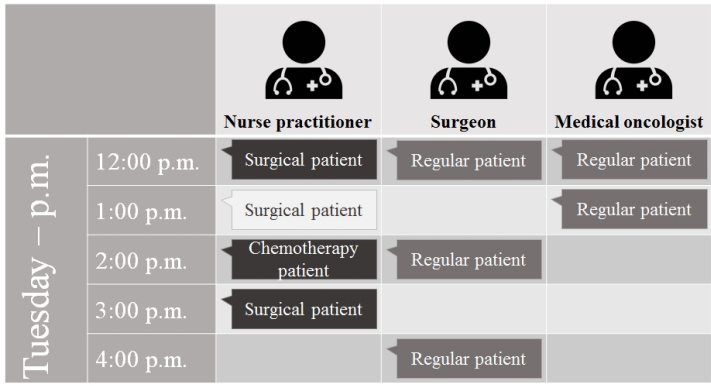
6.3 Literature

In the post-MTM clinic, patients need to visit multiple professionals in a fixed order. However, the specific appointment requirements are uncertain at the decision moment for the blueprint design. Therefore, we first evaluate the literature on multi-disciplinary scheduling in Section 6.3.1. Thereafter we discuss open access scheduling in more detail in Section 6.3.2, as the uncertainty in appointment requirements can be modelled as an open access system. In Section 6.3.3, we conclude by assessing the possibilities for multi-disciplinary scheduling with open access requirements.

6.3.1 Multi-disciplinary scheduling

In the literature review of Chapter 2, we saw that the problem at hand can be classified as a capacity planning problem of a flow-shop system with variable

Figure 6.2 Example of an NP schedule and clinicians' blueprint schedules of a small multi-disciplinary clinic with two clinicians – We consider consultations for multi-disciplinary patients with proposed surgical treatment or chemotherapy treatment, and regular consultations. The empty slots in the clinicians' schedules are available for multi-disciplinary patients on a FCFS basis.



appointment arrivals and appointment requirements. However, to the authors knowledge, no researchers have analyzed flow-shop systems with uncertain appointment requirements. As the simplified problem is already NP-hard, simulation techniques and heuristics are the most promising approaches to solve the problem.

Multiple authors consider the planning of flow-shop type multi-disciplinary systems, for example in oncology [178, 262] and primary care practices [222]. Liang et al. [178] analyze the impact of scheduling methods on the oncology clinic performance, where patients visit an oncologist and a nurse for chemotherapy treatment. Although various patient routings are considered, they consider this as given in their model. Saremi et al. [269] address the appointment scheduling of patients with various service sequences as well. They determine the appointment time of each patient in order to optimize a combination of waiting time and completion time. However, the number of patients per patient type, and thus the patient routing, are known in advance. Oh et al. [222] sequence patient appointments using a stochastic integer programming model with the sample average approximation approach. They included the effects of uncertainty in service time, but fixed the patient routing requirements. Romero et al. [262] use simulation to evaluate different appointment scenarios in which appointment blocks are reserved for multi-disciplinary patients. This way, they prove the feasibility of a one-stop-shop for basal cell carcinoma. However, they do not optimize the amount of capacity that needs to be reserved for serving the multi-disciplinary patients.

Simulation is the most widely applied technique in the literature studying the organization of multi-disciplinary scheduling. Simulation is used to analyze the performance of multiple clinics under a variety of scenarios, including various

appointment rules and appointment schedules [88, 156, 178, 188, 223, 262, 268, 269], which show significant improvements compared to the current practice in partnering health care centers.

Besides simulation, heuristics are applied. Both local search methods [266, 269] and other meta-heuristics [238] are applied to develop patient schedules, as well as approximate stochastic approaches [222] and simple planning rules [175, 263].

Since multi-disciplinary appointment scheduling involves multiple facilities that share patients, multiple performance indicators should be evaluated both for each facility at a local level as well as for the full system at a global level [196]. Not only the performance of the system, but also patient performance and clinician performance is taken into account in the literature.

Concluding, multi-disciplinary systems with precedence constraints are complex systems. Therefore, researchers focus on approximate solutions, such as simulation and heuristics, in order to optimize or evaluate the performance of these systems. To the author's knowledge, approaches to optimize or evaluate multi-disciplinary systems with stochastic patient routing are not available in the literature.

6.3.2 Open access scheduling

Open access scheduling is also known as same day scheduling, advanced access scheduling, short-notice scheduling, and walk-in scheduling [145, 249]. It entails the planning of multiple patient classes with different planning horizons. Open access approaches are introduced by Murray and Tantau [215], and were quickly adopted to reduce the effect of no-shows and cancellations [251, 261], as an alternative to overbooking strategies (e.g.,[94, 291]). Since this introduction, multiple researchers have researched the organization of open access appointment scheduling, both with a multi-day focus as well as an intra-day focus.

The multi-day focus concerns the percentage of appointment slots to reserve for open access patients [89, 248, 249, 290, 322], since this percentage influences among others the queue length and overtime [89]. Contrary to most available literature, Wiesche et al. [322] consider flexible capacity, to cope with varying patient arrival rates during the week in a primary care clinic. They use an integer programming approach to determine the optimal capacity, taking open access and regular patients into account, and evaluate the system performance by a stochastic simulation.

The intra-day focus concerns the sequence of fixed and open access appointments slots during the day [322], and the allocation of open access patients to appointment slots [21]. Since most authors evaluate multiple appointment sequences by a simulation study (e.g.,[158, 322]), only little work is performed on the optimization of these blueprint schedules. Peng et al. [233] optimize the number and position of open access and regular appointments by developing a blueprint schedule that minimizes the patient waiting time and clinician idle and overtime. Due to the high problem complexity of real life cases, their Genetic

Algorithm (GA) approach in combination with simulation requires high computational effort.

Few studies focus on the combination of multi-day and intra-day decisions. Kortbeek et al. [164] optimize the blueprint schedule considering both walk-in and scheduled patients, while allowing walk-in patients to be deferred. They develop two queuing models and propose a heuristic approach to generate appointment schedules based on these models.

Most open access literature consider a single-provider service system with fixed deterministic appointment intervals [261], where the capacity for each day is fixed and known [249], and where the demand and arrival rates are given [164, 248, 251]. The schedules of all providers involved are often assumed to be independent, both for providers of the same patient population, as well as for up- and downstream appointments [248, 249, 251]. Furthermore, a wide range of (combinations of) performance indicators is considered.

Concluding, open access scheduling focuses on determining the number of appointment slots and the sequencing of open access and fixed appointment slots. Where most literature assumes independent schedules, in our experience, schedules of clinicians influence each other, as the arrival rate at a walk-in clinic is determined by appointment schedules or service rates of upstream clinics. Open access scheduling with dependent schedules is an open question in the literature [249].

6.3.3 Contribution

Despite the long tradition of appointment planning in the literature, there has not been any attention for developing blueprint scheduling for multi-disciplinary appointment planning with open access requirements, as multi-disciplinary clinics are an emerging area in health care.

The blueprint schedule design of a multi-disciplinary clinic with open access requirements requires an integrated optimization approach, in which all appointment schedules are jointly optimized. To the authors' knowledge, the optimization of multiple clinics with open access requirements has not been considered before (as also argued by [39]). In addition, in all relevant literature, schedules of clinicians are assumed to be independent (see [248, 249, 251]), while in our experience with multiple hospitals, they depend on each other. Therefore, we analyze a multi-disciplinary clinic with open access requirements and dependencies between various clinicians tackling both open challenges.

Furthermore, we consider both multi-day and intra-day planning decisions. Since the sequencing of multi-disciplinary patients with the NPs influences the arrival rates at the clinician types, decisions on the available capacity for regular patients and on the slot sequencing of multi-disciplinary patients should preferably be made together.

Concluding, our contribution is threefold. First, we develop a model that includes dependent patient demand in open access models, an open challenge according to [249]. Second, multi-appointment planning in an open access context

is considered in this model, an open challenge according to [39]. Third, practical applications are presented in a case study of a real life health care setting.

6.4 Formal problem description and solution approach

To address the uncertainty in multi-disciplinary patient routing in the multi--disciplinary clinic, we adopt a stochastic programming approach. Stochastic programming has been applied in various health care settings. Min and Yih [205] developed a stochastic program to include uncertainty in surgery durations and length of stay for operating room scheduling, Bagheri et al. [18] used a stochastic programming approach for nurse scheduling, and Qu et al. [250] applied stochastic programming to appointment scheduling. To the best of our knowledge, stochastic programming has not been applied to multi-appointment planning with uncertain patient routing before. Section 6.4.1 formulates the problem as a Stochastic Integer Program (SIP). In Section 6.4.2 the recourse model is presented.

6.4.1 Problem formulation

Before we define the problem, we first introduce some notation, as summarized in Table 6.1. We use a set notation, where T are the time slots, and S the clinician types. The first clinician type ($s = 1$) corresponds with the NPs, who have a schedule in which appointments can be scheduled in time slots 1 to $|T| - 1$. The remaining clinician types s ($s \in S^*$) are the ones who have schedules in which slots can be pre-booked for regular patients, or are left empty for walk-ins from multi-disciplinary patients in time slots 2 to $|T|$. A clinician type has a capacity c_s, which means c_s clinicians of type s are available to see a patient per time slot.

The number of arrivals that will be referred to clinician type s follows a multinomial distribution. Since there is a finite number of possible arrival patterns, we can evaluate the performance of all possible scenarios, relative to their probability masses. For each of these arrival scenarios ξ, the probability of occurrence can therefore be calculated using the probability mass function of the multinomial distribution:

$$\phi^{\xi} = P(X_2^{\xi} = x_2^{\xi}, \text{ and ..., and } X_S^{\xi} = x_S^{\xi}) \frac{(Tc_1)!}{\prod_{s \in S^*} x_s^{\xi}!} \prod_{s \in S^*} P_s^{x_s^{\xi}}, \quad (6.1)$$

whereby the sum of all x_i should be equal to the total amount of appointment slots $c_1|T|$ of the NP. Note that for $|S| = 3$ this corresponds to the binomial distribution:

Table 6.1 Notation

Index and set	Definition
$t \in T, t \in \tilde{T}$	time slots that are in regular and overtime respectively, with $\tilde{T} = \{\|T\|+1, \|T\|+2, \ldots\}$, and $T^* = T \cup \tilde{T} \setminus \{1\}$
$s \in S$	clinician types, with $S^* = S \setminus \{1\}$
$\xi \in \Xi$	scenarios

Parameter	Definition
x_s^ξ	number of multi-disciplinary patients that arrive in scenario ξ and are referred to clinician type s
P_s	proportion of multi-disciplinary patients that will be referred to clinician type s for which $\sum_{s \in S^*} P_s = 1$ holds
c_s	capacity of clinician type s
ϕ^ξ	probability of scenario ξ, as derived from equation (6.1)
$\epsilon_1, \epsilon_2, \epsilon_3$	objective function weights

Variable	Definition
$X_{s,t}^\xi$	number of appointment slots reserved for multi-disciplinary patients in scenario ξ that will be referred to clinician type s in time slot t
$Y_{s,t}$	number of pre-booked appointment slots scheduled for clinician type s in time slot t
O_s^ξ	expected overtime in scenario ξ for clinician type s
W_s^ξ	total expected waiting time in scenario ξ for multi-disciplinary patients referred to clinician type s
$L_{s,t}^\xi$	queue length in scenario ξ at time t for clinician type s
I_s^ξ	idle time in scenario ξ for clinician type s

$$\phi^\xi = P(X_2^\xi = x_2^\xi \text{ and } X_3^\xi = x_3^\xi) = \frac{(Tc_1^\xi)!}{x_2^\xi! x_3^\xi!} P_2^{x_2^\xi} P_3^{x_3^\xi} = \binom{Tc_1}{x_2^\xi} P_2^{x_2^\xi} (1-P_2)^{Tc_1 - x_2^\xi}. \quad (6.2)$$

To optimize the blueprint schedule for all scenarios, we minimize for all clinicians $s \in S^*$ the expected overtime O_s^ξ, multi-disciplinary patient waiting time W_s^ξ, and the idle time I_s^ξ. To determine the waiting time, we also introduce the queue length $L_{s,t}^\xi$ for clinician type s in time slot t. Note that the queue length in overtime is determined as well, denoted by $L_{s,\tilde{t}}^\xi$. The weights for the overtime, waiting time, and idle time objectives are ϵ_1, ϵ_2, and ϵ_3 respectively.

In the stochastic program, all possible referral scenarios are to be evaluated. Therefore, we need two additional decision variables. $X^{\xi}_{s,t}$ is the number of appointments in the agenda of first clinician type ($s = 1$), that will be referred to clinician type s ($s \in S^*$), scheduled in time slot t in scenario ξ. $Y_{s,t}$ is the number of pre-booked appointments for regular patients for clinician type s ($s \in S^*$) in time slot t. This variable is independent of the scenarios, since it reflects the tactical level blueprint schedule, which has to be set before the realization of the patient arrivals.

The formal problem definition is as follows:

$$\min \sum_{\xi \in \Xi} \phi^{\xi} \left(\epsilon_1 \sum_{s \in S^*} O^{\xi}_s + \epsilon_2 \sum_{s \in S^*} W^{\xi}_s + \epsilon_3 \sum_{s \in S^*} I^{\xi}_s \right) \tag{6.3}$$

s.t.

$$\sum_{t \in T} X^{\xi}_{s,t} = x^{\xi}_s \quad \forall s \in S^*, \xi \in \Xi, \tag{6.4}$$

$$X^{\xi}_{1,t} = 0 \quad \forall t \geq |T|, \tag{6.5}$$

$$\sum_{s \in S^*} X^{\xi}_{s,t} = c_1 \quad \forall t \in T, \xi \in \Xi, \tag{6.6}$$

$$Y_{s,t} \leq c_s \quad \forall t \in T \setminus \{1\}, s \in S^*, \tag{6.7}$$

$$Y_{s,1} = c_s \quad \forall s \in S^*, \tag{6.8}$$

$$L^{\xi}_{s,t} \geq X^{\xi}_{s,t} + Y_{s,t} - c_s \quad \forall s \in S^*, \xi \in \Xi, t = 1, \tag{6.9}$$

$$L^{\xi}_{s,t} \geq L^{\xi}_{s,t-1} + X^{\xi}_{s,t} + Y_{s,t} - c_s \quad \forall t \in T^*, s \in S^*, \xi \in \Xi, \tag{6.10}$$

$$O^{\xi}_s \geq L^{\xi}_{s,|T|} \quad \forall s \in S^*, \xi \in \Xi, \tag{6.11}$$

$$W^{\xi}_s \geq \sum_{t \in T \cup \tilde{T}} L^{\xi}_{s,t} \quad \forall s \in S^*, \xi \in \Xi, \tag{6.12}$$

$$I^{\xi}_s \geq c_s |T| + O^{\xi}_s - \sum_{t \in T} \left(Y_{s,t} + X^{\xi}_{s,t} \right) \quad \forall s \in S^*, \xi \in \Xi, \tag{6.13}$$

$$\textit{all variables} \in \mathbb{Z}^+. \tag{6.14}$$

The objective is to minimize the weighted overtime, waiting time, and idle time, relative to the probability of each possible scenario of multi-disciplinary patient arrivals (6.3). For every scenario, the number of appointments to be scheduled for clinician type 1 (e.g., the NP) is given by the population distribution, and thus evaluated for every scenario (6.4). Note that the final slot of the booking horizon cannot be used by the NP (6.5). The number of these appointments should be equal to the capacity of this clinician type (6.6). Also, for the remaining clinician types, the number of pre-booked appointments cannot exceed the capacity (6.7). Note that in the first time slot of the booking horizon, no multi-disciplinary patients can be seen, thus all appointment slots can be filled

with pre-booked appointments (6.8). The queue length equals the queue length of the previous period plus the new arrivals (both walk-in and scheduled) minus the capacity of the clinician type (6.9)-(6.10). We assume only one clinician per clinician type to work in overtime, if necessary. Therefore, the number of overtime patients equals the number of overtime slots, which is equal to the queue length of the last time slot for each clinician type (6.11). Note that this equation can be replaced with (6.15) to include multiple clinicians serving overtime patients:

$$O_s^{\xi} \geq \sum_{t \in \tilde{T}} L_{s,t-1}^{\xi} \quad \forall s \in S^*, \xi \in \Xi. \tag{6.15}$$

The waiting time for each clinician type is the sum of all queues during the planning horizon, together with the waiting that occurs in overtime (6.12). Finally, the idle time equals the total time in which the clinicians of a clinician type are unoccupied during the planning horizon (6.13). All variables should be nonnegative (6.14).

The number of scenarios $|\Xi|$, grows with the number of appointment slots $c|T|$ and the number of clinician types $|S|$, following the multinomial distribution:

$$|\Xi| = \binom{c|T| + (|S| - 1) - 1}{(|S| - 1) - 1}. \tag{6.16}$$

For a small clinic instance, with 6 time slots with capacity 3, and 5 clinician types, 1,330 scenarios should be evaluated. For instances of clinics with 10 time slots with capacity 4, and 6 clinician types, 123,410 scenarios need to be evaluated. This shows that the problem becomes intractable for large instance sizes, through the high number of scenarios. Therefore, the Sample Average Approximation (SAA) approach will be applied, which approximates the objective function by considering a random selection of all possible scenarios [4, 159]. To apply the SAA approach, we reformulate the stochastic program as a recourse model in Section 6.4.2.

6.4.2 Recourse model

In the two-stage stochastic program with recourse, the first stage decides upon the optimal blueprint schedules for the clinician types. This decision is made at the tactical level, and is fixed for every possible scenario. In the second stage, the optimal scheduling strategy for the multi-disciplinary patients at the NP is determined by minimizing the recourse function, given the realization of multi--disciplinary patient arrivals.

Through the recourse formulation, as presented in Appendix I, it is seen that the recourse model is hard to solve, as it requires a high number of integer recourse functions to be solved. However, the constraint matrix that defines the feasible region of $X_{s,t}^{\xi}$ of the integer recourse function is *totally unimodular*, as all

determinants of the constraint matrix are 0, +1, or -1, and each column has two non-zero entries, which sum up to 0. Therefore, we can use the LP-relaxation to solve our integer program if the right-hand side is integer. Since all parameters and variables are integer, as c_s, $|T|$, and $Y_{s,t}$ are integers, this allows us to use a relaxation of $X^{\xi}_{s,t}$ as a continuous variable between 0 and 1.

6.5 Approximation algorithms

To find a solution to the problem, we first propose to solve the deterministic version of our problem in Section 6.5.1, by using expected values for all stochastic variables. This is the current practice in our partnering hospital and the literature. However, the stochastic nature of multi-disciplinary patient arrivals is not taken into account in this approach, which leads to solutions that are not robust in practice. Therefore, we apply the SAA approach in Section 6.5.2.

6.5.1 Average scenario

To be able to solve large instances of the mathematical program (6.3) - (6.14) from Section 6.4.1, we evaluate the deterministic version of the model, in which we assume the multi-disciplinary patient arrival rate to follow the (rounded) average scenario. This approach reflects the current hospital practice, where they designed the blueprint schedules based on the expected patient flow, rounded to the nearest integer. Furthermore, we think this approach also reflects the approach taken in the literature, where the patient case mix is assumed to be fixed, deterministic, and known.

The objective function of the original problem (6.3) is replaced by an easier evaluation (6.17), only considering one single scenario:

$$\text{minimize } \epsilon_1 \sum_{s \in S^*} O_s + \epsilon_2 \sum_{s \in S^*} W_s + \epsilon_3 \sum_{s \in S^*} I_s \tag{6.17}$$

This new model will provide a feasible solution to the original problem. Through the elimination of scenario evaluation, the complexity of the model is decreased, and therefore, the model can be evaluated within reasonable time. We assess the expected quality of the solution in reality, by simulating 1,000 realizations of the system, and evaluating the performance of these realizations.

6.5.2 Sample Average Approximation approach

The SAA approach approximates the objective value by evaluating a sample of $|N|$ scenarios. The scenarios in the sample are randomly drawn from the scenario population. The SAA approach does not only provide a solution, it also assesses the solution quality. Both lower and upper bounds to the objective of the stochastic program with corresponding optimality gap and confidence intervals

are provided [30]. We follow the SAA approach as proposed in [4, 30, 159], and refer to them for an in-depth description of this algorithm. The objective value as defined in (6.3), can be approximated by the average costs of all selected scenarios (6.18). This gives the following objective for a given solution $\hat{x}$:

$$\min \frac{1}{|N|} \sum_{n \in N} \left(\epsilon_1 \sum_{s \in S^*} O_s^n + \epsilon_2 \sum_{s \in S^*} W_s^n + \epsilon_3 \sum_{s \in S^*} I_s^n \right) \tag{6.18}$$

The constraints corresponding to the mathematical model of the SAA approach are constraints (6.31) - (6.40), as shown in the appendix to this chapter, where the full scenario set Ξ ($\xi \in \Xi$) is replaced by a sample set of scenarios N ($n \in N$).

This algorithm generates $|M|$ replications of $|N|$ samples for which the SAA model is solved. For each replication m, we generate a random sample of size $|N|$, and let $\hat{v}_{|N|}^m$ be the optimal objective value, and $\hat{x}_{|N|}^m$ be the corresponding optimal solution for replication m. When these values are computed for all replications, we evaluate statistical bounds over the total number of replications $|M|$. We have:

$$\bar{v}_{|N|}^{|M|} = \frac{1}{|M|} \sum_{m \in M} \hat{v}_{|N|}^m, \tag{6.19}$$

which is an estimator of the objective function $E[\hat{v}_{|N|}]$, and thus a lower bound to the optimal solution [159]. Furthermore, we have:

$$Var_{\bar{v}_{|N|}^{|M|}} = \frac{1}{|M|(|M|-1)} \sum_{m \in M} \left(\hat{v}_{|N|}^m - \bar{v}_{|N|}^{|M|} \right)^2, \tag{6.20}$$

which is an estimator of the variance of $E[\hat{v}_{|N|}]$.

Through the Central Limit Theorem, we can determine the 95% confidence interval ($\alpha = 0.05$) of the lower bound by:

$$\left[\bar{v}_{|N|}^{|M|} - \frac{z_{a/2} * \sigma_{\bar{v}_{|N|}^{|M|}}}{\sqrt{|N|}}, \quad \bar{v}_{|N|}^{|M|} + \frac{z_{a/2} * \sigma_{\bar{v}_{|N|}^{|M|}}}{\sqrt{|N|}} \right]. \tag{6.21}$$

Furthermore, an independent random sample of size $|N'|$ is generated. To compute the upper bound, the independent sample of size $|N'|$ is used to estimate the true objective value $\hat{g}_{|N'|}(\hat{x}_{|N|}^m)$, using (6.22), and the solution variance $Var_{\hat{g}_{|N'|}(\hat{x}_{|N|}^m)}$, using (6.23).

$$\hat{g}_{|N'|}(\bar{x}) = \frac{1}{|N'|} \sum_{n \in N'} \left(\epsilon_1 \sum_{s \in S^*} O_s^n + \epsilon_2 \sum_{s \in S^*} W_s^n + \epsilon_3 \sum_{s \in S^*} I_s^n \right) \tag{6.22}$$

$$Var_{\hat{g}_{|N'|}(\bar{x})} = \frac{1}{|N'|(|N'|-1)} \sum_{n \in N'} \left[\left(\epsilon_1 \sum_{s \in S^*} O_s^n + \epsilon_2 \sum_{s \in S^*} W_s^n + \epsilon_3 \sum_{s \in S^*} I_s^n \right) - \hat{g}_{|N'|}(\bar{x}) \right]^2 \tag{6.23}$$

We can determine the 95% confidence interval ($\alpha = 0.05$) of the upper bound by:

$$\left[\hat{g}_{|N'|}(\bar{x}) - \frac{z_{\alpha/2} * \sigma_{\hat{g}_{|N'|}(\bar{x})}}{\sqrt{|N'|}}, \quad \hat{g}_{|N'|}(\bar{x}) + \frac{z_{\alpha/2} * \sigma_{\hat{g}_{|N'|}(\bar{x})}}{\sqrt{|N'|}} \right]. \tag{6.24}$$

The optimality gap of each feasible solution $\hat{x}^m_{|N|}$ can now be estimated by subtracting the lower bound from the upper bound, $\hat{g}_{|N'|}(\hat{x}^m_{|N|}) - \bar{v}^{|M|}_{|N|}$, with corresponding estimated variance $Var_{\bar{v}^{|M|}_{|N|}} + Var_{\hat{g}_{|N'|}(\hat{x}^m_{|N|})}$. Furthermore, a final solution to the problem can be chosen from the replication sample, for example with the best value for $\hat{g}_{|N'|}(\hat{x}^m_{|N|})$.

6.6 Experiment design

This section describes the experiments. Section 6.6.1 describes the test instances and input parameters, and Section 6.6.2 presents the experiment results.

6.6.1 Input parameters

This section describes the input parameters and test instances, as summarized in Table 6.2.

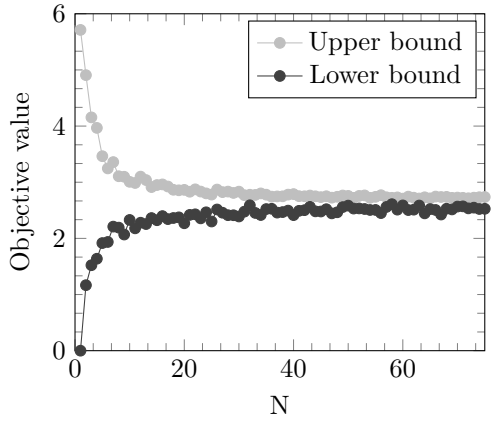
Input parameters We solve the SAA model for sample size $|N| = 25$, number of replications $|M| = 20$, and sample size to estimate the objective value $|N'| = 1,000$. The SAA approach is implemented in AIMMS 4 with CPLEX 12.6.

Test instances We consider an outpatient clinic with $|S| = 5$ clinician types. Since a clinic operates during the afternoon, in which typically a planning horizon between 8 and 10 time slots of 30 minutes is considered, we use a planning horizon of $|T| = 10$. Hereby we consider a capacity of $c = \{1, 2, 4\}$. This way, we vary over $c|T| = 10$, 20, or 40 appointment slots per outpatient clinic. 4 treatment specialists are considered, for which we vary over three scenario distributions. These distributions are given by $(P_2, \ldots, P_5) = (0.25, 0.25, 0.25, 0.25)$ for pattern 1, (0.1, 0.2, 0.3, 0.4) for pattern 2, and (0.4, 0.1, 0.1, 0.4) for pattern 3. Recall that multi-disciplinary patients cannot be referred to clinicians of type 1, since these clinicians diagnose the patient. Equal weights are assigned to ϵ_1, ϵ_2, and ϵ_3.

Table 6.2 Experiment settings

Parameter	Settings
$\|N\|$	25
$\|M\|$	20
$\|N'\|$	1,000
$\|S\|$	5
$\|T\|$	10
c	1,2,4
$(P_2, \ldots, P_5)$	(0.25, 0.25, 0.25, 0.25), (0.1, 0.2, 0.3, 0.4), (0.4, 0.1, 0.1, 0.4)
$\epsilon_1, \epsilon_2, \epsilon_3$	$\frac{1}{3}$

Figure 6.3 Objective value behavior with increasing number of scenarios $|N|$



6.6.2 Experiment results

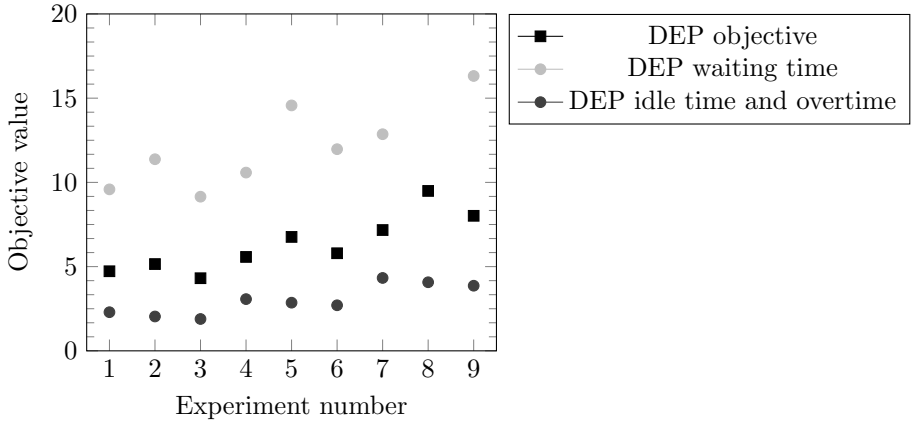
This section first describes the outcomes of the experiments to set the input parameters. Thereafter, the results of the different test instances are discussed.

Input parameter setting experiments The total number of possible scenarios follows from the number of clinician types and the number of appointment slots, as determined with equation (6.16). Unfortunately, the stochastic program becomes intractable if all possible scenarios are evaluated. Therefore, we determine a reasonable sample size in terms of solution quality and computation time.

To evaluate the amount of samples and replications, we analyze the scenario with Pattern 1 in more detail. Figure 6.3 shows the objective value behavior with different values for $|N|$. The objective value converges, and it can be seen that $|N| = 25$ samples will provide a reasonable optimality gap.

The solution quality increases with an increased number of samples and increased number of replications, against a price of computation time. For our problem instances, a sample size of $|N| = 25$ and replication number of $|M| = 20$ showed to give good solutions. In the remainder of this research, all experiments

Figure 6.4 True objective value behavior of the deterministic equivalent problem



are performed with $|N| = 25$ and $|M| = 20$, unless stated otherwise.

Experiment results of test instances Table 6.3 shows the results of the experiments. We analyzed the difference between the stochastic and deterministic approach, the effect of the clinic size, and the impact of different population distributions.

As Table 6.3 shows, the deterministic equivalent problem always derived an objective value of 0. Since only one scenario is evaluated, the schedules of the clinician types can be exactly adapted to the NPs' schedule. Thus, no waiting time, idle time, and overtime are incurred. However, as we can see in the evaluation of the deterministic equivalent solutions with 1,000 realizations, there will be an equal amount of overtime and idle time, as well as a large amount of waiting time in practice, which adds up to two to three times the performance of the more robust solution of the SAA approach (see Figure 6.4). Note that the idle time and overtime have equal values, as the deterministic equivalent solution fills all appointment slots. For each incurred idle appointment slot, a patient needs to be seen in overtime. Thus we can conclude that the SAA solution is more robust in practice, as it encounters for uncertainties in arrivals.

When the clinic size increases, the planning performance of the clinic slightly reduces, as can be seen from Table 6.3. Furthermore, the scenario distribution has impact on the schedule performance. Pattern 1 showed to have worse performance than patterns 2 and 3, which can be explained by the fact that every clinician's schedule has the same degree of uncertainty. In the other patterns, some clinician types get less referred multi-disciplinary patients, which means less disturbance by multi-disciplinary patients. On the other hand, some clinician types get more referred multi-disciplinary patients, which gives them economies of scale.

Table 6.3 Results of experiments

Exp. no.	Settings c	$\lvert T\rvert$	$\lvert cT\rvert$	p	SAA Obj.	Det. approach Obj.	True obj.
1	1	10	10	1	2.200	0	4.616
2	1	10	10	2	1.893	0	5.150
3	1	10	10	3	1.867	0	4.224
4	2	10	20	1	2.613	0	5.651
5	2	10	20	2	2.267	0	6.725
6	2	10	20	3	2.307	0	5.838
7	4	10	40	1	3.293	0	7.217
8	4	10	40	2	3.120	0	9.697
9	4	10	40	3	2.787	0	8.217

6.7 Case study

This section presents a case study of the hepato-pancreato-biliary (HPB) clinic of UMC Utrecht. In Section 6.7.1, UMC Utrecht's HPB clinic is described to give some context. Section 6.7.2 gives the input parameters and describes the case study instance. Finally, Section 6.7.3 presents the case study results using the SAA approach.

6.7.1 HPB department

UMC Utrecht's HPB cancer clinic provides care to patients with a (possible) tumor in their liver, pancreas, gallbladder, or biliary. In 2015, 318 new multi-disciplinary patients were seen in this clinic, which faces a growing patient demand. Every Tuesday, an MTM is conducted to assess all multi-disciplinary patients who were referred to UMC Utrecht, as well as the patients who need a second-opinion or patients who experienced recurring physical discomfort. Each patient has an intake (and possible additional diagnostic tests) in the morning of the same day. Four different medical specialties are present in the MTM meeting, in line with the possible treatment options: an oncological surgeon, a gastro intestinal physician, a radiotherapist, and a medical oncologist. Furthermore, the NP, pathologist, radiologist, genetic counselor, and some paramedical staff join the MTM.

During the afternoon, the multi-disciplinary clinic takes place, with consultation possibilities for all four specialties. Since surgery and chemotherapy are the most frequently recommended treatment modes, these specialties are present with multiple staff members. Furthermore, regular patients are seen by the four specialties for follow-up consultations, to ensure a high clinician occupation rate. These patients are pre-scheduled depending on the patient and clinician's preferences.

UMC Utrecht has provided real life data to evaluate the blueprint schedule design for the HPB clinic. The data spans the period of January 2015 to June

2016. Furthermore, the HPB oncology department of UMC Utrecht is exploring a growth scenario, in which is collaborated with multiple neighboring hospitals. Therefore, we evaluate this growth scenario as well.

6.7.2 HPB instances and input parameters

Current situation We consider a small outpatient clinic with $|S| = 5$ clinician types, with a planning horizon of $|T| = 8$, each consisting of 2 time slots of 30 minutes. Thus, the available capacity equals $c = 2$, which gives $|cT| = 16$ appointment slots.

The proportion of a population that requires a specific treatment modality, typically depends on the tumor types and the treatment possibilities per tumor type. Let $|A|$ be the number of tumor types, and let the k_a be the proportion of the population with this specific tumor type. The probability that a multi-disciplinary patient with tumor a gets treatment s is denoted with $p_{a,s}$ ($\sum_{s \in S^*} p_{a,s} = 1 \ \forall a \in A$). Therefore, through probability mapping, the proportion P_s of all appointments that will be referred to clinician type s can be determined by:

$$P_s = \sum_{a \in A} k_a p_{a,s} \quad \forall s \in S^*. \tag{6.25}$$

From the hospital data, we derived the population distribution and referral probabilities. The population distribution is given by $(k_1, \ldots, k_4) = (0.21, 0.10, 0.29, 0.40)$, and Table 6.4 gives the referral probabilities to the surgeon (surg.), oncologist (onc.), radiotherapist (RT), and gastro intestinal physician (GI). This gives a scenario distribution of $(P_2, \ldots, P_5) = (0.3208, 0.3113, 0.1849, 0.1830)$.

Since hospital staff was divided on the weights of the three performance measures, we evaluate various weight scenarios, as shown in Table 6.5.

Table 6.4 Referral probabilities from clinician type 1 to other clinician types per appointment type for the HPB case study

Appt. type	Surg.	Onc.	RT	GI
1	0.46	0.10	0.14	0.30
2	0.38	0.28	0.34	0.00
3	0.38	0.27	0.35	0.00
4	0.19	0.46	0.05	0.30

To compare the potential savings for UMC Utrecht using the results of the model, we also analyze the current way of working, which can be approximated by the deterministic equivalent of the stochastic problem.

Growth scenario As the current outpatient clinic size is rather small, the HPB departments of UMC Utrecht and its neighboring hospitals will merge into one multi-disciplinary HPB cancer clinic. In this new clinic, the same number of clinician types and appointment slots are considered ($|S| = 5$ and $|T| = 8$), but

Table 6.5 Weight settings, including overtime (ϵ_1), waiting time (ϵ_2), and idle time (ϵ_3)

Weight scenario no.	ϵ_1	ϵ_2	ϵ_3
1	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
2	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{1}{5}$
3	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{2}{5}$
4	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{5}{8}$

each clinician type has capacity $c = 4$. Furthermore, the population is expected to slightly change, to $(k_1, \ldots, k_4) = (0.25, 0.11, 0.25, 0.39)$. We assume the referral probabilities remain the same as in the current situation, which gives $(P_2, \ldots, P_5) = (0.3259, 0.3027, 0.1794, 0.1920)$.

Concluding, 12 case study experiments are executed, as shown in Table 6.6.

Table 6.6 Case study experiment design

Exp.no.	$\|S\|$	$\|T\|$	c	Population distribution	Weight scenario no.
CS1	5	8	2	$(0.3208, 0.3113, 0.1849, 0.1830)$	1
CS2	5	8	2	$(0.3208, 0.3113, 0.1849, 0.1830)$	2
CS3	5	8	2	$(0.3208, 0.3113, 0.1849, 0.1830)$	3
CS4	5	8	2	$(0.3208, 0.3113, 0.1849, 0.1830)$	4
DE5	5	8	2	$(0.3208, 0.3113, 0.1849, 0.1830)$	1
DE6	5	8	2	$(0.3208, 0.3113, 0.1849, 0.1830)$	2
DE7	5	8	2	$(0.3208, 0.3113, 0.1849, 0.1830)$	3
DE8	5	8	2	$(0.3208, 0.3113, 0.1849, 0.1830)$	4
CS9	5	8	4	$(0.3259, 0.3027, 0.1794, 0.1920)$	1
CS10	5	8	4	$(0.3259, 0.3027, 0.1794, 0.1920)$	2
CS11	5	8	4	$(0.3259, 0.3027, 0.1794, 0.1920)$	3
CS12	5	8	4	$(0.3259, 0.3027, 0.1794, 0.1920)$	4

6.7.3 Case study results

The results of the case study experiments are shown in Figure 6.5. Note that for the deterministic equivalent experiments, the true objective value is shown.

In both the current situation and the growth scenario, the SAA approach found good quality solutions for the different weight patterns, as the gap between the upper bound and lower bound is reasonably small. Better performance for specific performance indicators can be derived, depending on the weight settings.

However, a higher weight on specific indicators comes at a cost of reduced performance on the other indicators. More specifically, a trade-off has to be made between waiting and overtime, and the idle time, as solutions with better waiting and overtime often face worse idle time performance.

In the growth scenario more multi-disciplinary patients are seen by the clinicians, which reduces the uncertainty in their schedules. This is reflected in the lower SAA objective values for the growth scenario compared to the current situation relative to the size of the clinic (a difference of factor 1.48). However, through the higher amount of staff and patients, higher absolute total overtime, idle time, and waiting time are incurred.

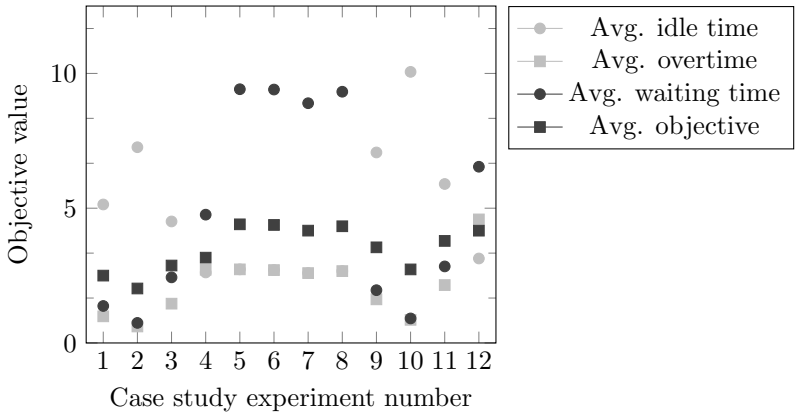
Compared with the current way of working, all proposed SAA solutions show better overall performance, of 50 minutes on average, which correspond with associated savings of 21% of the total clinic time. The performance on overtime and waiting time is improved for all weight settings, with up to 260 minutes less waiting time in total. However, in terms of idle time, the current situation might outperform the proposed SAA solutions. This is caused by the different amount of regular patients that are pre-booked in the the multi-disciplinary clinic. As can be seen from the figure, more patients are served in the current situation, as the overtime minus the idle time is greater than the overtime minus the idle time in the SAA solutions. However, the UMC Utrecht decision makers do not aim to serve as many regular patients as possible, but to serve all of their patients with as few waiting time and overtime as possible, given a reasonable idle time performance. The idle time performance is influenced by the amount of regular patients scheduled, the more regular patients, the less idle time. If idle time is not important at all, no regular patients will be scheduled, as this way the waiting and overtime are minimized. Therefore, the weight given to the idle time, includes the weight given for serving a high amount of patients. Note that the SAA solution for experiment 4 shows improved performance on all performance indicators, including the idle time, through the high weight on idle time. This shows that the SAA approach is capable of finding better overall schedules than the current way of working.

Concluding, we were able to find good schedules for the HPB clinic practice, based on various weight settings. The clinic has to decide which weight settings are important to them, as the overtime, idle time and waiting time measures vary according to specific settings. Furthermore, they have to make a final decision on which blueprint schedule to implement.

6.8 Conclusions and discussion

This chapter considers a two-stage stochastic program with integer recourse for the scheduling of multi-disciplinary cancer clinics. To solve this scheduling problem, an SAA approach is adopted. Experiments show and that the amount of uncertainty in patient arrivals influences the possible performance of a clinic, and that both for theory and practice good schedules can be obtained using this

Figure 6.5 Results of case study experiments



approach, which improves the current situation with 21% on average.

Health care practitioners should carefully discuss how to set the weights for waiting, idle, and overtime, as these affect the resulting schedules. In situations where clinics do not incorporate uncertainty in patient routing, and determine their schedule on the average patient mix, such as in UMC Utrecht's current situation, high weight is (unintentionally) put on idle time, as a high utilization is striven for. However, this might not reflect a clinic's intentions, which shows a thorough analysis of the current clinic's schedules is required.

Since all multi-disciplinary patients are discussed at the MTM, this research considers offline planning. In UMC Utrecht, the required treatment for all multi-disciplinary patients is known before the scheduling of multi-disciplinary patients in the NPs' agendas, as this scheduling step is done during the briefing preceding the clinic. In UMC Utrecht's practice, this situation therefore reflects reality. However, in a more general situation, one might want to schedule each multi-disciplinary patient at the time of their appointment request. This requires online planning, for which the stochastic model still can be used. The totally unimodular property needs to be dropped in this case, as the required treatment is not known at the time of the appointment request. It is left for future research to extend the current model to a multiple-stage stochastic program in which this new stochastic variable is taken into account.

This research was based on a few assumptions. First, we assumed a fixed slot structure for the blueprint of all clinicians. However, it is known that blueprint schedules without predefined slot structures might result in better performance [63]. Further research should show which slot structure is preferred for multi-disciplinary clinics.

Second, we considered a fixed clinic capacity, an unlimited demand of regular patients, and a fixed amount of multi-disciplinary patients. In practice, demand for multi-disciplinary care varies over the weeks. Hospitals tend to handle this varying demand in several ways. We chose to fix the capacity, and postpone

multi-disciplinary patients that arrive after all slots are filled to next week's clinic. Another way is to always accept multi-disciplinary patients that arrive, and serve them in overtime. In this case, one could adapt the number of appointment slots of the NPs in such a way, that it covers the maximum demand in for example 95% of the cases, and add an extra constraint to minimize the number of overbooked NP slots if possible.

Third, we assumed a fixed service duration for all patients. Although service duration variability has an impact on the performance of health care clinics (e.g, [222]), in our case study data on the amount of service duration variability was not known. Furthermore, as our model would explode when adding all sources of variability, we chose to incorporate uncertainty in patient routing over uncertainty in service duration, as the impact of a patient not visiting a provider is higher than the impact of a patient having a shorter visit with a provider. Further research is required to incorporate more sources of variability into one model, such as variability in patient arrivals, service durations, and capacity availability.

Fourth, the objective function of our model includes multi-disciplinary patient waiting time, and clinician idle time and overtime. We chose to not take patient access time into account, for multi-disciplinary as well as regular patients. Since all multi-disciplinary patients are assumed to already be present in the hospital, all multi-disciplinary patients have equal arrival times. Including the access time for multi-disciplinary patients would therefore not influence the optimal solution. Furthermore, the access time of regular patients is influenced by factors outside the system under review, as regular patients are also served in other clinics. Therefore, the access time for regular patients cannot be accurately determined.

Fifth, the model assumes that referrals can only be done to clinicians of other types. In health care settings, it might be the case that the clinician who gives the diagnosis, is also one of the treating clinicians. For example the surgeon or the gastro intestinal physician can have this double function in both the diagnostic as well as the treatment phase. Further research should be done to analyze the effect of recurrent referrals.

Sixth, we assumed patients are served on a FCFS basis. However, in a clinic environment it is debatable whether FCFS is the most equitable priority rule for patients, as patients have diverse priorities, due dates, and appointment series. Furthermore, it is questionable whether it is necessary to use FCFS, as long as patients are served within a reasonable time. As we analyzed a multi-disciplinary clinic with patients with two sequential appointments, the FCFS priority rule is feasible. In a multi-disciplinary clinic with varying numbers of appointments (e.g., patients that can have 2, 3, or 4 appointments in a row), other priority rules, for example based on the expected remaining throughput time, might be more suitable.

Seventh, we analyzed the multi-disciplinary clinic independent from the morning processes. Incorporating the effect of appointments in the morning into the afternoon schedule, or jointly optimizing the morning and afternoon clinics might give improved results which necessitates multi-appointment scheduling solutions with three or more appointments.

Based on our research, several directions for implementation in practice are present. First, the blueprint schedule solution can be implemented, which shows schedulers in which appointment slots a regular patient can be scheduled, and which appointment slots should be left empty, to encounter for multi-disciplinary patients. Second, the second stage model can be used to plan the multi-disciplinary patients in the agenda of the NPs, after the MTM. Currently, UMC Utrecht implemented a new blueprint schedule and use simple planning rules for real-time scheduling based on the results of this research.

Since the patient population of a hospital changes over time, and since new treatment modalities can be introduced, the model should be used by hospitals in a dynamic way. We advise hospital managers to redesign their blueprint schedules at least once a year. Our integrated optimization approach, in which all appointment schedules are jointly optimized, can help hospital managers to efficiently organize their multi-disciplinary care systems.

6.9 Appendix I

The stochastic problem of equations (6.3) - (6.14) can be formulated as the following recourse model:

$$\min E[Q(x,\xi)], \tag{6.26}$$

s.t.

$$Y_{s,t} \leq c_s \quad \forall t \in T \setminus \{1\}, s \in S^*, \tag{6.27}$$
$$Y_{s,1} = c_s \quad \forall s \in S^*, \tag{6.28}$$
$$Y_{s,t} \in \mathbb{Z}^+ \quad \forall t \in T, s \in S^*, \tag{6.29}$$

where $E[Q(x,\xi)]$ is the corresponding recourse function, with:

$$Q(x,\xi) = \min \epsilon_1 \sum_{s \in S^*} O_s^\xi + \epsilon_2 \sum_{s \in S^*} W_s^\xi + \epsilon_3 \sum_{s \in S^*} I_s^\xi, \tag{6.30}$$

s.t.

$$\sum_{t \in T} X_{s,t}^\xi = x_s^\xi \quad \forall s \in S^*, \tag{6.31}$$
$$X_{1,t}^\xi = 0 \quad \forall t \geq |T|, \tag{6.32}$$
$$\sum_{s \in S^*} X_{s,t}^\xi = c_1 \quad \forall t \in T, \tag{6.33}$$
$$\tag{6.34}$$

$$L^{\xi}_{s,t} \geq X^{\xi}_{s,t} + Y_{s,t} - c_s \quad \forall s \in S^*, t = 1, \tag{6.35}$$

$$L^{\xi}_{s,t} \geq L^{\xi}_{s,t-1} + X^{\xi}_{s,t} + Y_{s,t} - c_s \quad \forall t \in T^*, s \in S^*, \tag{6.36}$$

$$O^{\xi}_{s} \geq L^{\xi}_{s,|T|} \quad \forall s \in S^*, \tag{6.37}$$

$$W^{\xi}_{s} \geq \sum_{t \in T} L^{\xi}_{s,t} + \sum_{\tilde{t} \in \tilde{T}} L^{\xi}_{s,\tilde{t}} \quad \forall s \in S^*, \tag{6.38}$$

$$I^{\xi}_{s} \geq c_s|T| + O^{\xi}_{s} - \sum_{t \in T} \left(Y_{s,t} + X^{\xi}_{s,t} \right) \quad \forall s \in S^*, \tag{6.39}$$

$$\textit{all variables} \in \mathbb{Z}^{+}. \tag{6.40}$$

CHAPTER 7

Simulating the multi-disciplinary outpatient clinic

7.1 Introduction

The organization of multi-disciplinary clinics is challenging, as processes in multiple departments have to be jointly optimized. Furthermore, this has to be done according to the performance indicators of many stakeholders. Chapter 6 discusses the design of blueprint schedules for multi-disciplinary clinics, a decision on the tactical level of control. Using these blueprints, planners can assign patients to specific slots and control the daily operations. This operational level capacity-to-patient assignment problem is the focus of this chapter.

The organization of cancer treatment becomes more complex, involving multiple specialties, as patients receive more personalized care and medical practitioners specialize. Many cancer patients in UMC Utrecht receive (neo)adjuvant therapies, for example through radiotherapy or chemotherapy. In Chapter 6 we studied a multi-disciplinary outpatient clinic in which patients receive their diagnosis and information about their cancer treatment at the same day. However, when multiple specialists are involved in the treatment of the patient, it would be beneficial for the patient to also meet with the corresponding care providers of the additional therapies as well, to be fully aware of the treatment strategy of choice. In this chapter we therefore extend the multi-disciplinary clinic of Chapter 6 by incorporating patient types with more than 2 consultations. Note that the order in which these consultations take place is not fixed, as it does not matter which treating clinician is seen first.

The analysis in this chapter supports the operational planning decisions in two of UMC Utrecht's multi-disciplinary outpatient clinics in which patients subsequently visit two or more care professionals. In such an outpatient clinic, in the current situation, a medical assistant is present to align the patients and provider schedules. This medical assistant receives a list of patients and the required consultations after the multi-disciplinary team meeting (MTM), and has to determine which patient is the next in line to visit each practitioner in the clinic following the MTM. He or she has to ensure that the waiting time for patients is minimized, and that the overtime of practitioners is minimized as well, which are two important performance measures in appointment planning. However, as

the assistant is not educated in advanced planning, scheduling, or logistics, the question arises what general and easily applicable planning rules result in a well performing clinic with low waiting and overtime rates. Hereby, both the routing as well as the prioritization of patients should be taken into account.

In addition to finding the best planning rules, UMC Utrecht would like to send each patient an invitation with the (approximate) start time of the first consultation in the multi-disciplinary clinic. This is challenging for two reasons. First, there is uncertainty in the number of patient arrivals, as it is not known in advance how many patients are in need of a multi-disciplinary approach each week. Second, the number of required appointments and the required providers for each patient that requests an invitation for a multi-disciplinary approach are not known at the moment of sending the invitation. Therefore, the clinic management, and the assistant in particular, wants to know the optimal invitation strategy, such that the moment of physical arrival of patients at the clinic ensures minimal waiting times during the course of the clinic, but also ensures efficient clinic operations.

To evaluate various planning rules for practical use in an environment with multiple stochastic elements, we develop a discrete event simulation (DES) model of the multi-disciplinary clinic to analyze the planning strategies. Computer simulation is a frequently used evaluation methodology for appointment planning and scheduling problems in healthcare, as shown by the large amount of literature reviews on healthcare simulation modeling (e.g., [43, 100, 118, 152, 217]). Furthermore, for multi-disciplinary planning on an online operational level of control it is the most frequently applied method, as shown in Chapter 2. The most relevant system studied is that of [153]. In their analysis of a multi-disciplinary oncology clinic, the authors state that dependencies between the care providers of the different disciplines can be captured in a DES model, and that coordinated clinic schedules are of advantage to both the patient as well as the provider. They considered the oncology clinic from a higher level, focusing on the relation between the MTM meeting with the morning and afternoon clinic, instead of a more in-depth focus on the required appointments within the afternoon clinic, which is the focus of our research.

In this study, we aim to prospectively assess the effect of multiple planning rules and invitation policies, in order to find the rules that will provide a good performance in practice. As clinic characteristics are of influence to the optimal process design of these clinics, the model is generic, and easily extendable to other multi-disciplinary situations. We show the applicability of this model to two case studies of UMC Utrecht.

This chapter is organized as follows: Section 7.2 introduces the simulation model, after which Section 7.3 presents the experiment design and results. Section 7.5 gives the conclusions, discussion, and opportunities for further research.

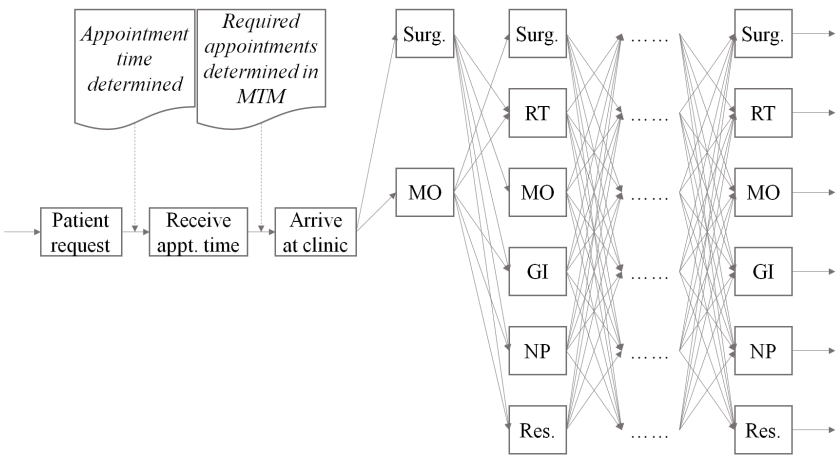
7.2 Simulation model

To fulfill the aim of this study, finding planning rules that balance the patients' waiting times and clinic efficiency, we develop a DES model, following the methodology of Law [171].

7.2.1 Process flow

Patients that require treatment based on a multi-disciplinary approach, for example because there are multiple treatment options or combinations of treatment modalities available, are referred to the multi-disciplinary clinic by hospital clinicians, or by clinicians from a referring hospital. After this referral, the medical assistant of the multi-disciplinary clinic invites the patient to come to the clinic the day on which the next MTM takes place. On this invitation, a specific time during that day is mentioned. Typically, patients already received a confirmation of their diagnosis, and only the staging of the tumor and a treatment plan are unknown. For this, clinicians have to discuss the patient in their MTM. The result of the MTM is a treatment plan for the patient, including the required appointments that are needed for the patient in the clinic. After the MTM is finished, the specialists go to the consultation rooms, and start the clinic session. They see all required patients, whom all have varying appointment requirements, in an order that the medical assistant determines. Patients arrive to the clinic's waiting area by the time they are invited through the invitation letter, and are subsequently seen by specialists according to the requirements from the MTM. When a patient finishes a consultation, and has another required consultation, he or she has to wait in the waiting area again, otherwise the patient can go home. The process flow from a patient perspective is visualized in Figure 7.1.

Figure 7.1 Patient process flow (Surg.=surgeon, RT=radiotherapist, MO=medical oncologist, GI=gastro intestinal physician, NP=nurse practitioner, Res.=researcher)



7.2.2 Input data

To simulate this process, several data inputs are required, with respect to the clinic, specialists, and patients. Two years of data (2015-2016) is derived from the hospital information system, together with estimations based on expert opinions.

Clinic data. We consider two cases from two UMC Utrecht clinics. Both clinics operate one day a week: Tuesday (referred to as Clinic 1) or Wednesday (referred to as Clinic 2). The MTM preceding the clinics starts at 12:30 PM, and takes 2 hours. Both clinics start at 2:30 PM, and end at 5:00 PM. If necessary, patients are seen in overtime to ensure all patients receive the care they need.

Specialist data. Based on the hospital data, there is a fixed number of specialists of a certain specialist type available during the clinic hours. We consider 6 specialist types: surgeons, radiotherapists, medical oncologists, gastro intestinal physicians, nurse practitioners, and researchers.

The average appointment duration μ_s for each of the specialist types s is estimated based on expert opinions, and does not depend on the patient type. In the simulation model, the appointment duration for each appointment follows a normal distribution, from which negative durations are excluded.

Table 7.1 displays the number of available specialists per type and the service times per specialist type.

Table 7.1 Specialist type characteristics

	Clinic 1			Clinic 2		
Specialist type	# specialists	μ_s	σ_s	# specialists	μ_s	σ_s
Surgeon	3	30	10	2	30	10
Radiotherapist	1	30	10	2	30	10
Medical oncologist	2	30	10	2	30	10
Gastro intestinal physician	2	30	10	2	30	10
Nurse practitioner	2	30	10	2	30	10
Researcher	2	30	40	2	30	40

Patient data. Patients arrive to the clinic following a Poisson distribution, that shows to be a good fit to the hospital data. A patient $p_{c,t}$ that arrives to clinic c is considered of a certain type t, based on the patient characteristics, which are known before the MTM. In this chapter, we consider patient types based on tumor type. Clinic 1 serves three patient types: pancreas patients, who account for 47% of the total population, liver patients, who account for 35% of the total population, and gal-bladder and biliary tract patients, which account for 18% of the population. Clinic 2 serves two patient types: esophageal patients (80%) and stomach patients (20%). The patient type determines from which selection of treatment plans the required appointments are drawn, as each patient

type has corresponding probabilities of the various treatment strategies, as shown in Figure 7.2 and 7.3. In the MTM meeting, treatment plans are allocated, depending on the patient type and tumor specific care pathways. Table 7.2 shows the possible treatment plans per patient type, together with their probabilities (derived from historical data). Note that there are no precedence relations between the appointments, except for the first appointment, which always has to be an appointment with the surgeon. In the case of a treatment plan without any surgeon involved, the first appointment is with the medical oncologist.

Figure 7.2 Fraction of patients with a certain number of appointments

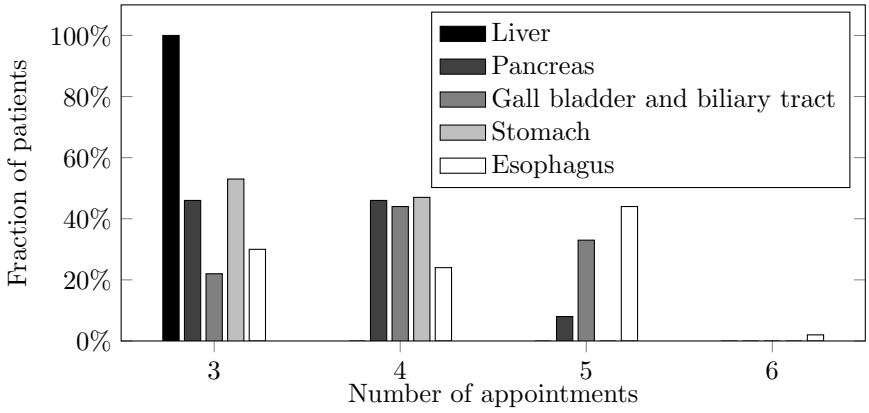
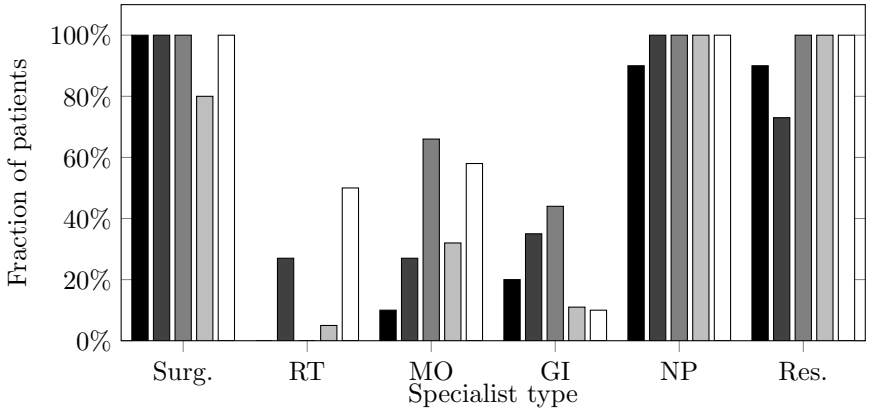


Figure 7.3 Fraction of patients that require consultation with each specialist type



7.2.3 Performance indicators

We use the simulation model to test the effects of several planning rules, thereby analyzing the following performance indicators for each intervention:

Table 7.2 Treatment plan characteristics per patient type

Clinic	Patient type	TP No.[a]	Surg.[b]	RT[c]	MO[d]	GI[e]	NP[f]	Res.[g]	Probability
1	liver	1	x				x	x	0.7
1	liver	2	x			x	x		0.2
1	liver	3	x		x			x	0.1
1	pancreas	1	x				x	x	0.46
1	pancreas	2	x		x	x	x	x	0.08
1	pancreas	3	x	x		x	x		0.27
1	pancreas	4	x		x		x	x	0.19
1	gall	1	x				x	x	0.22
1	gall	2	x		x	x	x	x	0.33
1	gall	3	x			x	x	x	0.11
1	gall	4	x		x		x	x	0.33
2	stomach	1	x				x	x	0.53
2	stomach	2			x		x	x	0.21
2	stomach	3	x	x			x	x	0.05
2	stomach	4	x			x	x	x	0.11
2	stomach	5	x		x		x	x	0.11
2	esophagus	1	x				x	x	0.30
2	esophagus	2	x	x	x	x	x	x	0.02
2	esophagus	3	x	x		x	x	x	0.02
2	esophagus	4	x	x	x		x	x	0.42
2	esophagus	5	x	x			x	x	0.04
2	esophagus	6	x			x	x	x	0.06
2	esophagus	7	x		x		x	x	0.14

[a] TP No. = Treatment plan number, [b] Surg. = surgeon, [c] RT = radiotherapist, [d] MO = medical oncologist, [e] GI = gastro intestinal physician, [f] NP = nurse practitioner, [g] Res. = researcher.

Waiting time: The time a patient physically spends in the waiting room. This includes the time before the first appointment, as well as the time between subsequent appointments, both in regular hours and in overtime. Thus, if two patients have waited 15 and 45 minutes respectively, the waiting time is 30 minutes on average.

Overtime: The sum of the time outside regular hours that patients are served. Thus, if two patients both had to stay 15 and 45 minutes respectively outside regular hours to be served, the overtime is 1 hour. Note that this measure penalizes overtime not from a provider perspective, by considering all providers separately through the service time of the patients, and also from a patient perspective, by including the overtime per patient instead of focusing on the total makespan.

Utilization: The average total consultation time of a specialist within regular opening hours as a percentage of the total regular opening hours. The utilization per provider is of specific interest, besides the average utilization, as the patient pathways show high variation in the demand for specialist types. Therefore, we expect the utilization to highly vary among specialist types.

As the proposed system is a new system design which is only operational since April 1st 2017, validating the model by comparing its output to the performance of the existing real life system is not yet possible. Therefore, we validated the model through extensive checks with subject-matter experts, such as clinicians, medical assistants, and clinic management. This way, we concluded that the model is a valid representation of reality.

7.3 Experiment design

In this section we describe the experiment design, whereafter we present the results in Section 7.4.

We use the simulation model to simulate various planning strategies for the operational control of the multi-disciplinary clinics. We experiment with the invitation strategy (i.e., what is the scheduled arrival time for each patient at the clinic?), the routing rules (i.e., what is the best appointment sequence for each patient?), and the prioritization rules (i.e., which patient from the pool of patients in the waiting room to see first?). All evaluated strategies are determined together with clinic staff, in order to increase the probability of implementation.

Invitation strategy. After a patient is referred to one of the multi-disciplinary clinics, an invitation letter is sent to the patient in which the appointment time of the first appointment is stated. As the course of the treatment is not yet known, no doctor can be assigned to the patient at this time. Patient invitations can be sent in an online or offline fashion, e.g., directly after the request is received, or by batching several requests and processing them simultaneously respectively. We explore three settings:

- *Online (On)*, in which immediately after the referral is received, a patient gets assigned an appointment time.
- *Daily (Day)*, in which at 6 PM each day, all referred patients of that day get assigned an appointment time.
- *Offline (Off)*, in which at 6 PM the day before the clinic takes place, all referred patients get assigned an appointment time .

In each of these settings, patients can be invited for their first appointment based on various planning rules. We explore four settings:

- *Single appointment time*, in which each patient is assigned the same appointment time, e.g., 3 PM, the moment that the clinic opens.
- *Equal spread*, in which each patient is assigned an appointment slot such that the arrival of patients is equally spread over the clinic hours.
- *Diverge patient type*, in which patients are scheduled based on their patient type. Herein, patients of a patient type with the highest expected number of appointments are scheduled first.
- *Alternate patient type*, in which patients of various patient types are alternately scheduled if possible.

If we consider appointment slots, we assume slots to be of the same length as the expected duration of the surgeons' consultations, who are the first clinicians to meet with the patients, with a fixed amount of 3 slots. If based on the planning rule, multiple patients classify for the same appointment slot, for example in the alternate patient type setting, the patient that arrived first is scheduled for the earliest appointment time. Note that in an online invitation setting, the diverge patient type and alternate patient type planning rules result in equal first come first serve (FCFS) scheduling practice, as each patient is scheduled individually. Similarly, for the single appointment time policy, the online, daily and offline settings will give equal results. Note that we do not consider advanced planning rules, which for example take possible future arrivals into account, as from a practical perspective these planning rules are hard to implement in practice. Therefore, incorporating future arrivals in decision making is an area of future research.

Routing rules. Each patient has a list with required consultations that should be finished before leaving the clinic. There are only few precedence constraints between those consultations, which gives flexibility in the routing of patients through the clinic. After a consultation with a clinician, a patient is referred to the next waiting room by the medical assistant. We explore three settings for the decision which appointment is next, and thus to which waiting area the patient is referred:

- *Fixed order (Fixed)*, the patient is referred based on the preferred appointment sequence by the clinicians.
- *Expected waiting time (Wait)*, the patient is referred to the specialist type with the lowest expected waiting time.
- *Idle providers (Idle)*, the patient is referred to the specialist type with the highest number of idle providers.
- *Random (Rnd)*, the patient is randomly assigned to a next specialist type.

If based on the routing rules multiple referrals are weighted equally, the next appointment is randomly assigned from this selection to a patient.

Prioritization rules. The selection of a new patient to consult from the patients that are waiting in the waiting room of a certain specialist type can be done based on various prioritization rules. We explore the following rules:

- *First come first serve (FCFS)*, in which the patient that arrived the first to the waiting area of this specialist type is seen first.
- *Last come first serve (LCFS)*, in which the patient that arrived the last to the waiting area of this specialist type is seen first.
- *Most remaining turnaround time first (MRTT)*, in which the patient with the highest expected remaining turnaround time is seen first.
- *Least remaining appointments first (LRTT)*, in which the patient with the lowest expected remaining turnaround time is seen first.
- *Highest turnaround time first (HighTT)*, in which the patient with the highest total turnaround time is seen first. This includes all previous appointments.
- *Lowest turnaround time first (LowTT)*, in which the patient with the lowest total turnaround time is seen first. This includes all previous appointments.
- *Random (RND)*, in which a random patient is selected from the waiting area.

If based on the prioritization rules multiple patients are weighted equally, one of these patients is randomly selected.

Considering all combinations of invitation, routing and prioritization strategies, we conduct a total of 252 experiments. We implemented the simulation model and the experiments in TechnoMatix Plant Simulation 11. For each experiment, we simulate 214 replications of one week, as we have a terminating system.

7.4 Results

The simulation experiments show that the best performing configuration depends on the performance indicator of interest. Figures 7.4 and 7.5 show the Pareto efficient frontier plots for average waiting time and average overtime of the experiments. Both of these figures show three performance clusters, which are characterized by the invitation rules. The upper clusters corresponds with the single appointment slot rule, the clusters on the left with the equal spread rule, and the middle clusters with the diverge and alternate patient type invitation rules. This shows that the invitation rules have the largest impact on the performance indicators of interest. Furthermore, it can be noticed from these graphs that within all clusters specific routing rules are on the efficient frontier. Each of the prioritization rules however, is prevalent on the efficient frontier. We will analyze these observations in more detail further below.

As the hospital strives for a robust approach, we not only evaluate the average performance, but also the performance at the 75 percentile. The latter means that in 75% of the time that the clinic operates, the performance will be within

Figure 7.4 Pareto frontier for Clinic 1

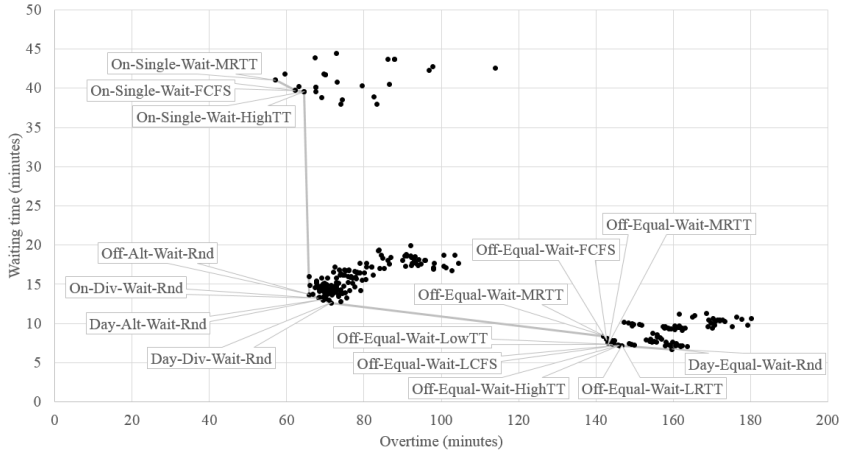
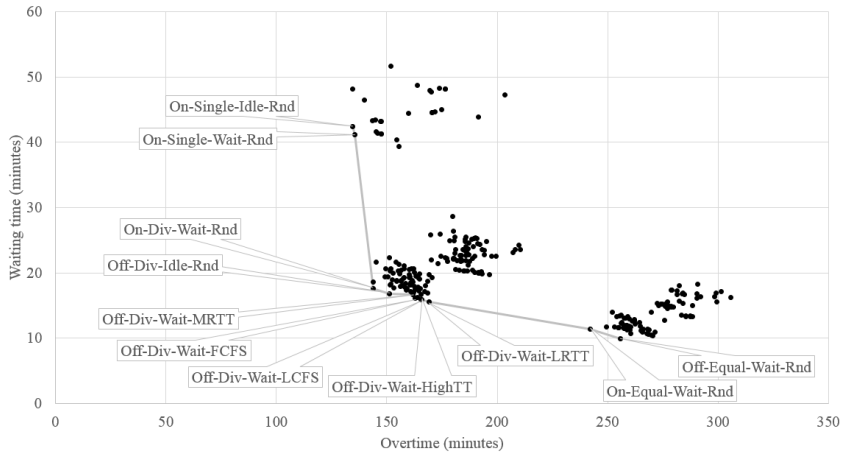


Figure 7.5 Pareto frontier for Clinic 2



that boundary. This way, less robust approaches, despite having a good average performance, are less favored. The best configurations for each of the performance indicators for Clinic 1, based on the 75 percentile, are:

– For waiting time: The configuration in which patients are invited in an online setting, on an equal spread basis, routed based on the expected waiting time, and prioritized based on any priority rule (excluding random prioritization).

– For overtime: The configuration in which patients are invited in an online setting, on a single appointment time basis, routed based on the expected waiting time, and prioritized based on the most appointments first. Note that the routing based on idle providers is slightly better on average, but has

a higher variation.

- For utilization: The configuration in which patients are invited in a daily setting, on an alternated patient type basis, routed based on the expected waiting time, and randomly prioritized. Note that the same configuration with an offline invitation strategy performs similarly, which is also the best strategy according to the weighted KPI setting.
- For weighted normalized KPIs: The configuration in which patients are invited in an offline setting, on an alternated patient type basis, routed based on the expected waiting time, and randomly prioritized.

The best configurations for each of the performance indicators for Clinic 2, based on the 75 percentile, are:

- For waiting time: The configuration in which patients are invited in an offline setting, on a equal spread basis, routed based on expected waiting time, and randomly prioritized.
- For overtime: The configuration in which patients are invited in an online setting, on a single appointment time, routed based on idle providers, and randomly prioritized.
- For utilization: The configuration in which patients are invited in an online setting, on a single appointment time, routed based on expected waiting time, and prioritized based on the lowest turnaround time first.
- For weighted normalized KPIs: The configuration in which patients are invited in an online setting, on a diverge patient type basis, routed based on expected waiting time, and randomly prioritized.

The results for each of these configurations are shown in Figures 7.6, 7.7, and 7.8. The white bars represent the best performing solutions for Clinic 1, and the gray bars for Clinic 2.

Figure 7.6 Waiting time performance for best configurations

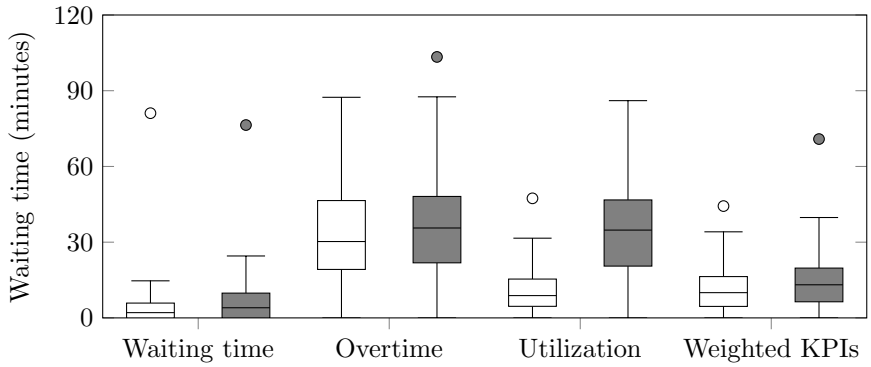


Figure 7.7 Overtime performance for best configurations

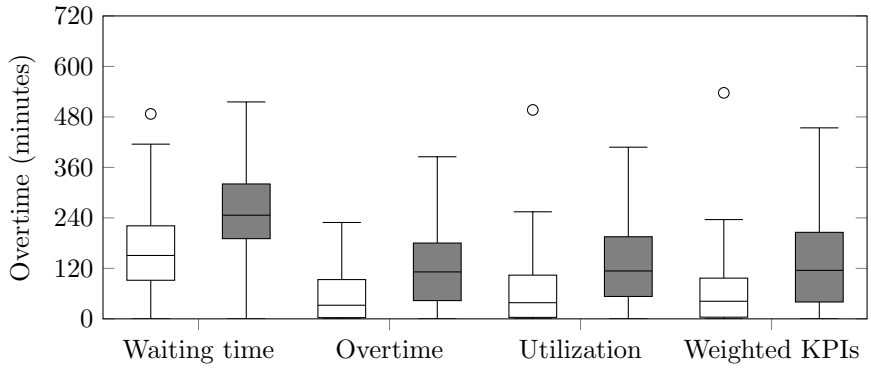
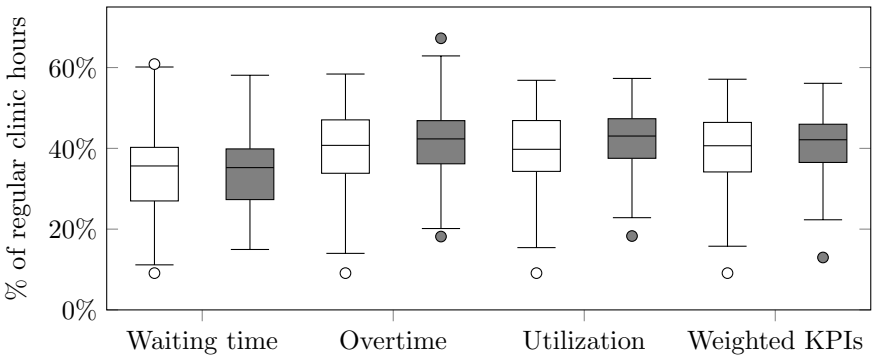


Figure 7.8 Utilization performance for best configurations



The average waiting time per patient varies highly over the reviewed configurations, with a maximum of 44.5 and 51.7 minutes on average per patient, and a minimum of 6.7 and 9.7 minutes on average per patient for Clinic 1 and 2 respectively. Where for Clinic 1 the NPs are the specialists that patients have to wait the most for on average, for Clinic 2 these are the surgeons and radiotherapists. This can be explained by the high demand for radiotherapists of Clinic 2 patients, and by the approximately similar demand for surgeons that Clinic 2 faces in comparison to Clinic 1, whereas they have a lower capacity. Through the higher waiting times for the surgeons, the outflow to the NPs and researchers is more leveled, and thus less waiting occurs at those specialists, although they face similar demand. The two clinics give a good example how solving one bottleneck (in this case the surgeons at Clinic 2) by adding more capacity, will result in another bottleneck (the NPs and researchers).

As can be seen in Figures 7.4, 7.5, and 7.7, both clinics experience overtime, even in the best performing configurations. This is as expected, as the clinic hours are limited. When patients arrive that require 4 or 5 appointments, the only way to see them in regular time, is to invite them right at the start of the clinic. However, as one does not know in advance which patients require that

many appointments, in practice, these patients arrive later during the day as well, inevitably resulting in overtime. A possible solution to solve this problem is to extend the clinic hours for certain specialist types (for example by starting one hour later, and finishing one hour later). Another solution is to invite all patients from the start of the clinic, as proposed in the single appointment time invitation strategy. However, this solution results in high patient waiting times (see Figure 7.9).

To derive a good utilization or weighted performance in Clinic 1, the best prioritization rule is the random strategy. In the remainder of this chapter we will show the low impact of various prioritization rules, which explains this unexpected result. The utilization of both the clinics is about 40% on average over all specialist types. When solely focusing on the utilization, this indicates there is room for increased demand. However, the overtime figures show otherwise. When looking into the specifics of the utilization by analyzing the utilization rates of the various specialist types individually, we see that in Clinic 1 the radiotherapist has a low average utilization, of 9% to 15%, as expected considering the treatment opportunities. Similarly, the medical oncologists and the gastro intestinal physicians experience low utilization rates (16% and 24% on average respectively). On the other hand, the surgeons, nurse practitioners and researchers all have a higher utilization, of around 50%. Note that through the precedence relation of the first appointment, surgeons can see most of their patients in regular time, whereas the NPs and researchers regularly have to wait during the start of the clinic, but finish their work in overtime, which heavily impacts the utilization figure of these specialist types. If all patients would have been seen in regular time for the NP and researcher, their utilization would be 75% and 63% on average respectively. In Clinic 2, a similar situation is present (expected utilization within working hours of NPs as well as researchers of 79%).

When we compare Clinic 1 with Clinic 2, we see that Clinic 2 faces higher demand, which can be seen through the higher expected overtime in combination with the slightly higher utilization measures. Although the same amount of patients is expected, this can be explained by the higher expected number of appointments that the patients of Clinic 2 require. To analyze the impact of higher and lower demand on the clinics, we evaluated the performance of both clinics in the best experiment settings. In a higher demand situation (+12.5%) the performance of Clinic 1 is similar to the performance of Clinic 2 in the current system.

In the remainder of this section, each of the experiment factors is analyzed in comparison to this best performing configurations. We highlight several specific experiment settings; the full results are available with the author upon request.

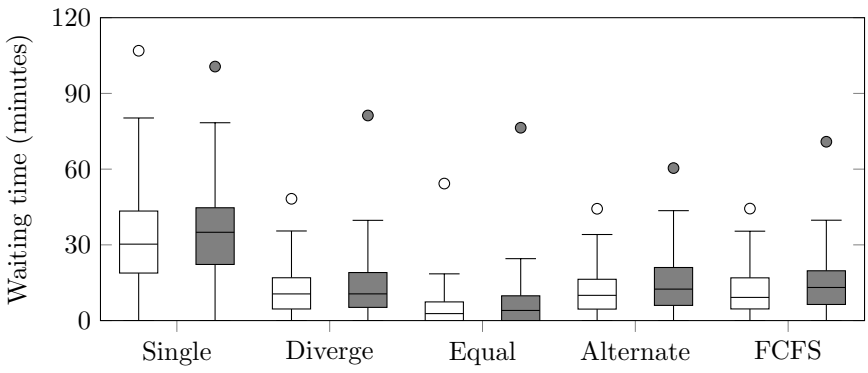
Invitation strategy. Figures 7.9, 7.10 and 7.11 display the performance of the experiments with regard to the various invitation strategies at an offline level, as well as an online FCFS strategy. The white bars represent Clinic 1, and the gray bars Clinic 2.

The figures show that the invitation strategy has a large impact on the perfor-

mance of the clinics. It is for example beneficial to equally spread patients over the appointment slots when waiting time is of importance, as the equal spread invitation strategy leads to an up to 32 minutes waiting time improvement for Clinic 1 ($p < 0.001$), and up to 31 minutes for Clinic 2 ($p < 0.001$). The alternate and diverge patient type invitation rules perform, and are also significantly better with 23-26 minutes on average compared to the single appointment time rule ($p < 0.001$). However, although the best in terms of waiting time, the equal spread invitation strategy is not preferred if highly valuing overtime performance, as it leads to significantly more overtime (e.g., 84 to 87 minutes more overtime on average for Clinic 1, and 100 to 109 minutes for Clinic 2, with $p < 0.001$, compared to all other considered rules in an online setting). The single appointment time rule performs best ($p < 0.001$). This is as expected, as overtime is minimized when all patients are invited as early as possible. The overtime behavior is also representative for the utilization performance, as the utilization plots show the inverse behavior of the overtime plots.

The moment of inviting the patients does not heavily influence the performance of the clinic. Although in general flexibility, which is created with the offline planning rule, is considered to be favorable, as better performing schedules can be derived through more knowledge about the demand, in this situation the benefits are small. A possible reason is that the extra information is limited, as the required appointments of a patient are not known when inviting the patient to the clinic.

Figure 7.9 Waiting time performance for invitation strategies



Routing rules. Figures 7.12, 7.13, and 7.14 display the performance of the experiments with regard to the various routing rules. The white bars represent Clinic 1, and the gray bars Clinic 2.

Despite the small benefits that can be seen in these figures, the routing rules based on expected waiting time or idle providers significantly outperform the fixed order and random assignment rules with up to 6 minutes less waiting time, and 30 minutes less overtime ($p < 0.001$). Distributing patients over (idle) providers

Figure 7.10 Overtime performance for invitation strategies

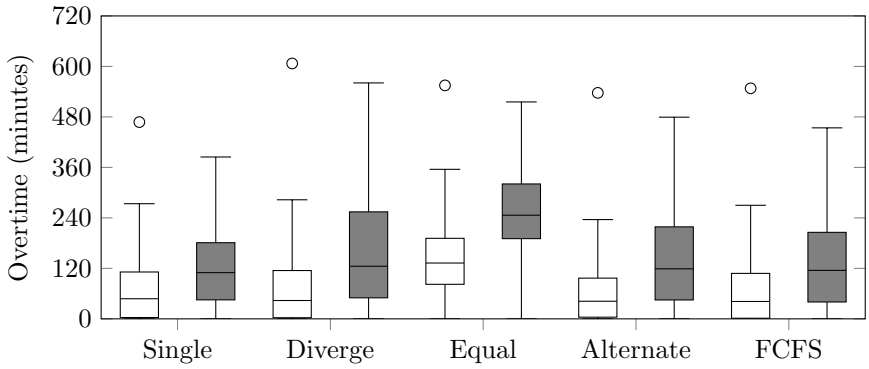
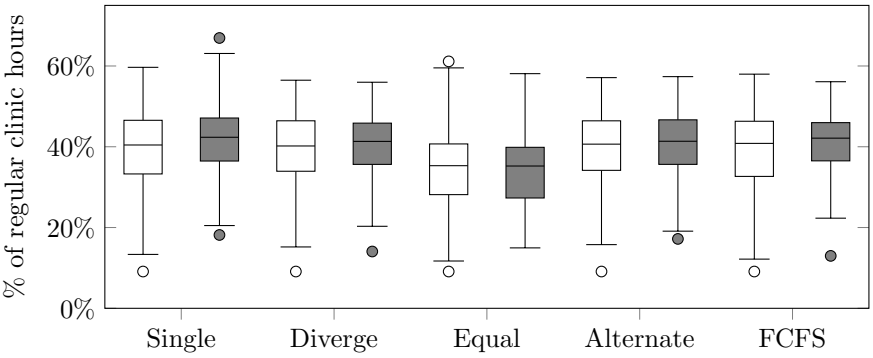


Figure 7.11 Utilization performance for invitation strategies



or providers with low expected waiting times, ensures that patients do not have to wait too long, and that providers do not stay idle too long, which decreases the probability of treatment in overtime, and increases the utilization in regular time. Comparing the performance of the expected waiting time rule and the idle providers rule shows that the expected waiting time rule results in significantly less waiting time (49 seconds for Clinic 1 and 86 seconds for Clinic 2, $p < 0.001$), but higher overtime (76 seconds for Clinic 1, $p < 0.001$, and 87 seconds for Clinic 2, $p = 0.004$). Note that although these effects are significant, the difference is negligible small for practice.

Prioritization rules. Figures 7.15, 7.16, and 7.17 display the performance of the experiments with regard to the various prioritization rules. The white bars represent Clinic 1, and the gray bars Clinic 2.

It is shown that the impact of varying the prioritization rule is rather low, as all rules show similar performance, and the best performing rule on one indicator is the worst rule on another indicator. For example, the best waiting time performance rule (HighTT) performs 1 minute (for Clinic 1) and 2 minutes (for Clinic 2) better on average compared to the worst performing rule (LRTT), and the

Figure 7.12 Waiting time performance for routing strategies

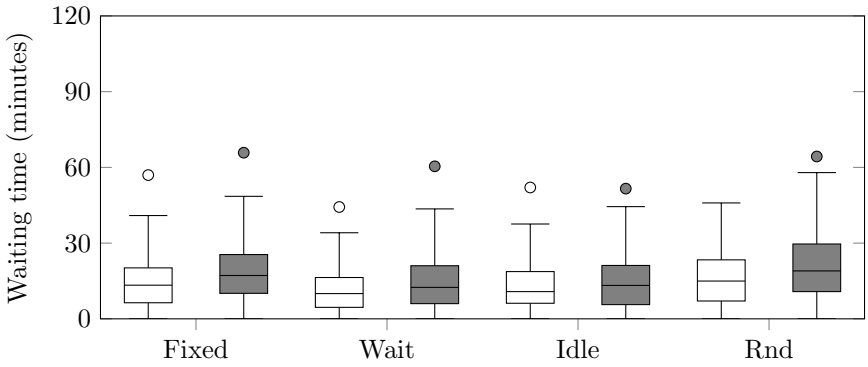
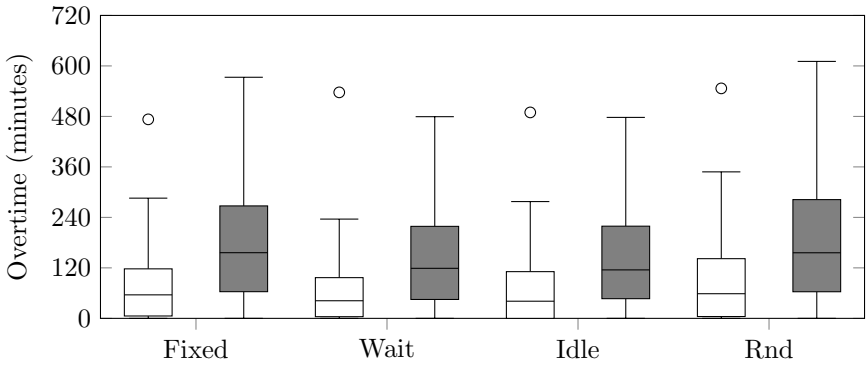
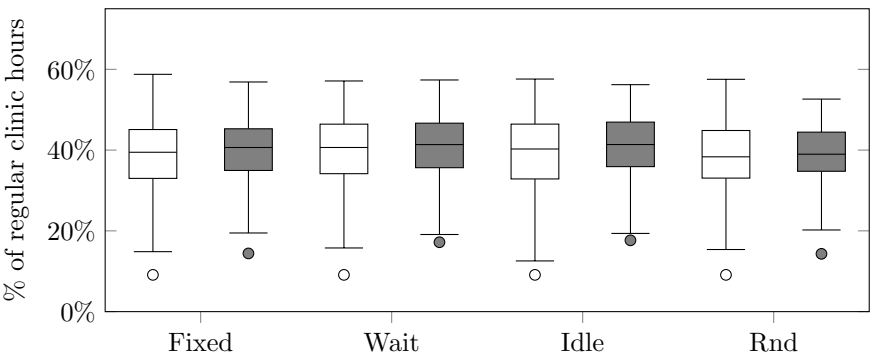


Figure 7.13 Overtime performance for routing strategies



best overtime performance rule (MRTT) performs 3 minutes (for Clinic 1) and 13 minutes (for Clinic 2) better on average compared to the worst performing rule (HighTT). A possible reason for this small difference is the impact of routing rules on the number of patients in the waiting area for a specialist type. Even the

Figure 7.14 Utilization performance for routing strategies



least favorable routing rules distribute patients over the specialists. Therefore, the specialists do not have to choose between patients that often, which makes the impact of prioritization rules low.

Figure 7.15 Waiting time performance for prioritization strategies

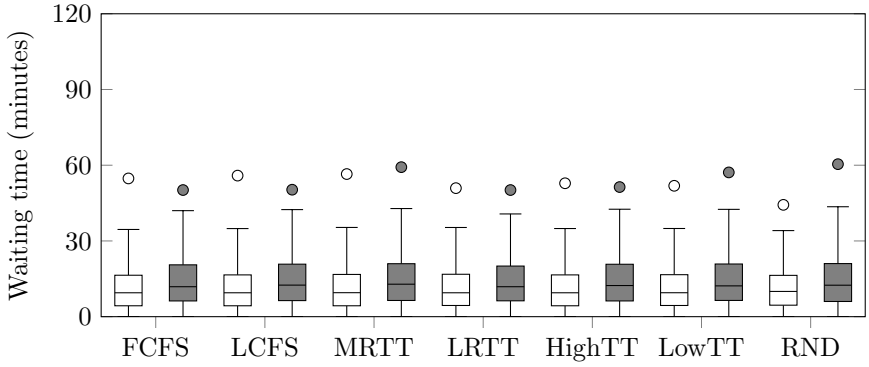
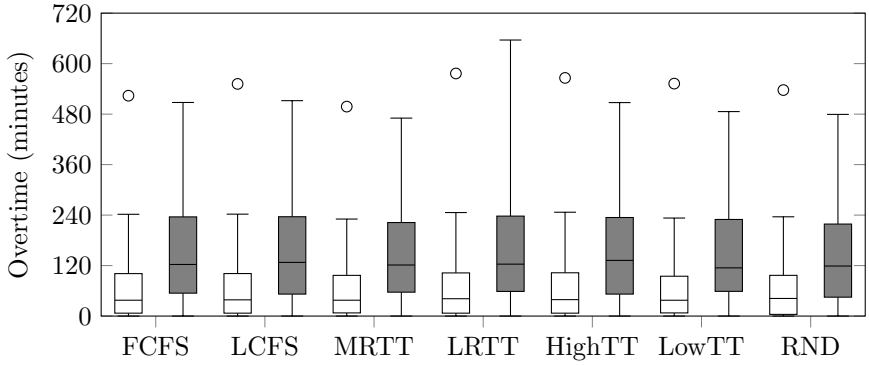


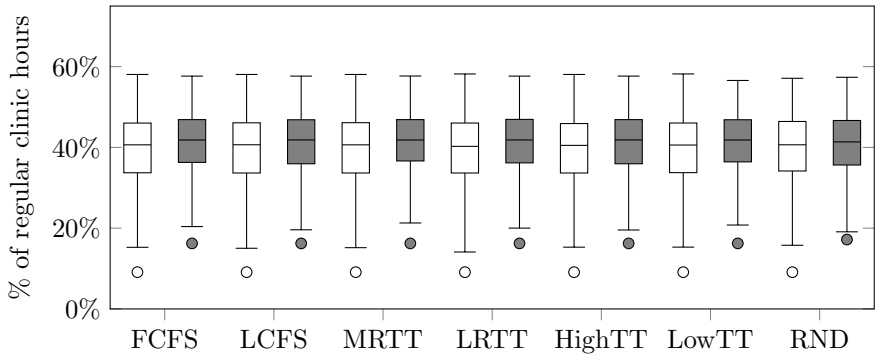
Figure 7.16 Overtime performance for prioritization strategies



7.5 Conclusions and discussion

In the Netherlands, there is an increase in outpatient clinics that provide a multi-disciplinary approach to derive treatment plans for cancer patients, together with providing the patients all necessary information that they need. The organization of such clinics is complex, as patient pathways are uncertain, multiple appointments are required, and multiple specialties are involved. In this chapter, we analyzed the organization of multi-disciplinary clinics, and investigated multiple invitation strategies, routing rules, and prioritization rules, in order to minimize the patient waiting time and provider overtime, and to maximize the system's utilization.

Figure 7.17 Utilization performance for prioritization strategies



The results show that decisions made the earliest have the largest impact on the clinics' performances. The patient invitation strategy enables the clinics to focus on the waiting time or overtime, or a combination of both. After this decision is made, smaller benefits can be gained by choosing the routing rule. The prioritization rules show the least impact.

The use of an invitation strategy shows to significantly improve the clinics' performance, either by using the equal spread strategy when focusing solely on the waiting time, by using a single appointment time when focusing solely on overtime, or by using the diverge or alternate patient type strategy considering a mix of performance indicators. We advise the clinic to choose the moment of sending these patient invitations based on patient preferences, instead of logistical outcomes, as the results of the simulation study do not show significant benefits of one strategy over the other. As an offline strategy is beneficial when incorporating knowledge about the demand in the scheduling strategy, further research can be done in exploring advanced invitation strategies incorporating knowledge about (future) demand.

Routing patients based on expected waiting time or the number of idle providers shows to perform significantly better than routing based on a fixed order or in a random manner. As for management assistants it is currently hard to adequately assess the expected waiting time for each of the waiting areas, we advise to route patients based on the number of idle providers. When no provider is idle, the management assistants can either randomly assign a patient to the next specialist type, or make an educated guess about which specialist type has the lowest expected waiting time, based on their experience with the clinic.

The prioritization rules do not heavily impact the clinics' performance, which offers room for the clinics to prioritize their patients in a way that suits their needs or that is perceived best by their patients.

The multi-disciplinary clinic faces challenges with the utilization of their specialists, which were confirmed by this study. For specific specialist types, the utilization is very low (down to 9%), whereas other specialist types experience a high utilization (up to 80%). One solution is to allow low utilized specialist types

to invite other patients to the clinic, for example using the approach of Chapter 6. However, this will generate more delays for the multi-disciplinary patients, as well as an increased overtime for all specialist types. Therefore, in making this decision, specialist types should not only consider the benefits for themselves (higher utilization), but also the side-effects for patients and colleagues from other specialties. A follow-up study, in line with Chapter 6, can analyze the effects of inviting other patients to specific timeslots during the multi-disciplinary clinic.

For the case study clinics, we identified challenges with the performance indicators through the setup of the clinic. For example, through the precedence relations that require the surgeon as the first consulted specialist type for most of the treatment plans, most of the other providers are not able to see any patients at the start of the clinic, resulting in idle time and overtime. Furthermore, patients may require up to 5 appointments of 30 minutes in 2.5 hours. This is only possible within regular hours when a patient is scheduled for the first appointment slot, otherwise the patient has to be seen in overtime. However, the number of appointments is not known when scheduling the patients. A possible solution is to let specific specialist types, such as the surgeons, start their consultations earlier, for example 1 hour earlier. This way, at the official starting time of the clinic, the other specialist types can immediately start seeing patients. This provides more room for patients with many appointments, and may enable them to finish on time as well. Further research is required to analyze the best starting times of each specialist type in relation to the possible care pathways of the patients and the invitation strategy, in order to ensure that the benefits with respect to overtime and utilization do not result in excessive waiting times.

When striving for a good waiting time time performance, the clinics have to determine how much waiting time they consider acceptable. Herein, targets need to be set for the overall waiting time in the clinic, as well as waiting times before each of the specialist types. A patient who waited 30 minutes in total, could have seen 3 specialists for which he waited 10 minutes each, or could have waited 30 minutes for one specialist. Note that the first case might be acceptable, whereas the second case should be prevented.

The design of a new multi-disciplinary cancer clinic in a multi-appointment and multi-provider setup, enables the hospital to reconsider the old planning processes, and to incorporate smart planning decisions from the first moment the clinic becomes operational. This research supports this opportunity. However, as the multi-disciplinary cancer clinics in UMC Utrecht are only operational since April 2017, we were unable to validate our model against real practice. Therefore, we suggest to compare the results from the simulation model with the first results from the actual clinics, to increase the validity of the model. Note that a thorough analysis of the data is required, as there might be flawed data because of starting errors in the first operational months.

For the same reason, we included appointment duration data based on expert opinions in the simulation model. Further research is required to derive good estimators of the duration parameters, for example through a data analysis of the

actual duration of the consultations after the clinic was implemented in practice.

In this chapter, we focused on the evaluation of planning and scheduling rules. In line with [153], we found DES to be suitable for modeling complex operational characteristics of the multi-disciplinary system. The use of a simulation methodology, such as DES, is especially powerful for creating support among hospital employees. Not only can the processes be visualized, which makes it easy for practitioners to understand what is happening, but it also enables interventions of staff members to be evaluated.

To incorporate optimal planning decisions into the experiment design of the simulation model, simulation optimization approaches can be explored to determine optimal routing in the clinic, given the expected arrivals of patients and the required appointments they need. The effect of various invitation strategies and prioritization rules can be further explored given the optimal patient routing through the clinic.

PART 4

treatment

Why Wait?
Organizing Integrated Processes in Cancer Care

Chapter 8

A.G. Leeftink and E.W. Hans. Case mix classification and a benchmark set for surgery scheduling. *Journal of Scheduling*, https://doi.org/10.1007/s10951-017-0539-8, 2017.

CHAPTER 8

Case mix classification and a benchmark set for surgery scheduling

8.1 Introduction

The application of Operations Management/Operations Research (OM/OR) to healthcare has been studied since the 1950s, and has gained particular attention over the past decade [145]. For elective patients, many key healthcare resources (such as outpatient clinics, diagnostic facilities, and operating rooms) are organized on an appointment basis. Therefore, many studies have looked at the scheduling of appointments or surgical procedures [42, 53, 56]. Brailsford et al. [43] concluded that the extent of actual implementation of the outcomes of healthcare simulation and modeling studies is disappointingly low, and has always been so. Startlingly few studies report evidence of implementation, although a relatively large proportion do demonstrate a conceptualized model [43]. The fact that real-world data is hard to obtain in healthcare may have contributed to Brailsford's conclusion. This unavailability is caused by privacy considerations, and by the simple fact that data registration is primarily done for medical and financial purposes. Although we observe some change, historically no need was felt to record data for operations management purposes. Nevertheless, the application of operations research models and computer simulation inherently requires a lot of data. Therefore, researchers predominantly resort to using theoretical and computer-generated data for their experiments, or use the limited amount of available real-world data, supplemented with computer-generated data. However, the applicability of these outcomes in other settings is questionable. The questions arise: How complex are these instances? How do these instances compare to instances from other hospitals (and perhaps in other countries)? How does the algorithm perform on such other instances? Finally, how relevant are the presented results for the scientific community, or even for a specific healthcare manager from another hospital?

Benchmark instances are ideal for comparing solution approaches for specific and well-defined problems. A well-known example is the Solomon instances set for benchmarking algorithms for vehicle routing problems (VRP) [280]. However, in healthcare scheduling, no widely used benchmark instances exist other than for the nurse scheduling problem [298]. Van Riet and Demeulemeester [259] underline

in a recent operating room planning review that test instances are needed that cover realistic hospital settings in order to align the operating room planning research.

Since many healthcare scheduling problems concern assigning patients to shared resources, these problems have a common denominator: scheduling a set of patients with a (stochastic) resource demand. This similarity allows for the creation of benchmark sets, which can be used by a wide variety of algorithms. If specific additional data is needed within a particular application setting, for example additional patient properties such as urgency, users can add such aspects themselves. Having standard benchmark sets available (1) helps to deal with a lack of data availability; (2) allows benchmarking algorithms for similar or identical problem types; and (3) allows for comparing obtained real-life instances to standard benchmark instances and for their classification.

One of the most studied topics in healthcare scheduling literature is surgery scheduling, also known as operating room planning and scheduling [53]. However, benchmarking the performance of surgery scheduling between different hospitals is difficult, as case mixes differ between such hospitals. A case mix describes the volume and characterization of all surgery types. Cardoen et al. [52] presented a classification scheme for classifying the surgery scheduling problem. However, the impact of the composition of the case mix on the performance of a surgery scheduling algorithm, which is instance-dependent, was not taken into account.

This chapter contributes to the literature in multiple ways: We propose a case mix classification scheme, which can also be used to typify and visualize surgery scheduling instances. Furthermore, we illustrate the application of the classification scheme to surgery scheduling datasets obtained from both academic and non-academic hospitals in the Netherlands, and some cases from the literature. We provide a benchmark set for the surgery scheduling problem, which is based on real-life data from Dutch hospitals, as well as theoretical data, and which may serve as a reference benchmark set for testing surgery scheduling algorithms. To ensure that the generated benchmark instances are sufficiently diverse, we introduce the concept of instance proximity, which is a measure for the similarity of instances. We introduce an instance proximity maximization approach for the instance generation procedure. Finally, we provide applications with which unlimited additional instances and samples can be generated, using statistical distributions based on real-life and theoretical data.

The remainder of this chapter is organized as follows. Section 8.2 discusses the literature about benchmarking, instance generation, and instance classification. In Section 8.3, we describe the problem definition and present the case mix classification. Section 8.4 applies the classification to real-life datasets and datasets from the literature. Section 8.5 presents the instance generator, and introduces the instance proximity maximization concept. Finally, in Section 8.6 we present our conclusions.

8.2 Literature

The classification of scheduling instances, and the development and use of benchmark sets, has been an important topic in the operations research literature [163]. To develop ideas for how to classify patient scheduling instances, we investigate the literature about benchmark sets and instance generation for planning and scheduling problems in Section 8.2.1. In Section 8.2.2 we discuss the conditions for a benchmark set to be effective.

8.2.1 Benchmark sets

A benchmark set is defined as a collection of instances of a class of combinatorial optimization problems [162]. A benchmark set is also referred to as an instance library. A well-known and widely used benchmark set for the VRP problem was provided by Solomon [280]. Researchers use this benchmark set and adapt the instances to their specific needs, by adding their own characteristics. A widely used extension of the Solomon benchmark set was developed by Gehring and Homberger [109], who added residual groups to the existing instances. Kok et al. [160] used the Solomon instances set and added time-dependent travel times and driving regulations to make them feasible for common, practical situations. More recently, Pillac et al. [240] presented a set of technician scheduling problem instances, extended from the Solomon instances set, by adding technician crews.

Benchmark sets also exist (among others) for project scheduling problems [161, 318], timetabling problems [247], and personnel scheduling problems [216].

A project scheduling problem library (PSPLib) was presented by Kolisch and Sprecher [161], which among others exists of instances generated by the project scheduling instance generator ProGen [91, 162]. In this library, several benchmark sets are available for researchers, who can use the benchmark set of their needs. These benchmark sets are updated over the years, depending on the progress in the project scheduling research field [161]. A combination of instances from PSPLib, added with release dates and global resources, were used as instances for the MISTA 2013 challenge [318]. Also, a combination of resource constrained project scheduling problem instances was used to generate multi-project scheduling instances, which are known as MPSPLib [141].

In their literature review on timetabling, Qu et al. [247] analyzed and characterized the available timetabling benchmark sets. They summarized the applied techniques and corresponding results on the benchmark sets. They concluded that benchmark sets should be updated and extended according to the needs of the research area and derived from real life data, in order to minimize the gap between theory and practice.

Multiple benchmark sets exist for personnel scheduling. The University of Nottingham presents an overview of various personnel scheduling benchmark sets, derived from researchers and from industry [78]. Musliu et al. [216] also provided personnel scheduling instances, which were generated by selecting a solution and randomly generate an instance based on this solution. This way,

they already have a (near) optimal solution for each of their generated instances.

Although various benchmark sets exist for well-known combinatorial optimization problems, for healthcare planning and scheduling there only exist various benchmark sets for nurse scheduling [46, 216, 299] and patient admission scheduling [35]. Typically, these benchmark sets, such as NSPLib [299], are not real-world-based instances but are generated randomly. Within the nurse scheduling research field, two competitions have been organized, for various problem configurations, such as multi-stage nurse rostering [62]. In the first competition, three tracks were presented, based on the available running time of the algorithms, including small, medium, and large sized instances to solve [135]. A patient admission scheduling benchmark set [35] has been generated on the basis of a single real-life instance from Demeester et al. [83]. Since only limited real life data was available, Ceschia and Schaerf [61] presented a benchmark set for patient admission scheduling, together with an instance generator, solution validator, and first solutions. The instances are generated based on randomly generated theoretical case mixes.

To the best of our knowledge, no widely used benchmark sets have been reported for other healthcare scheduling problems, such as the surgery scheduling problem [260]. Reasons for this may be the lack of data due to reluctance of hospitals to disclose data, the numerous different representations of essentially the same problem, and the many ways of evaluating solution procedures [298]. In the remainder of this chapter, we focus on generating and providing benchmark instances for the most studied healthcare problem, namely the surgery scheduling problem. We refer to Denton et al. [87] for an extensive problem description of the deterministic and stochastic surgery scheduling problem in multiple operating rooms.

8.2.2 Conditions for benchmark sets

An algorithm may perform well on one instance, but can have a poor performance on another comparable instance. To provide a benchmark set that represents real-world problems, but also allow for an instance-independent comparison of the performance of algorithms, benchmark sets need to systematically integrate the characteristics of the problem as their parameters [162]. According to Vanhoucke and Maenhout [298], a benchmark set should satisfy four conditions: diversity, realism, size, and extendibility. First, the instances of a benchmark set should be as diverse as possible, to facilitate unbiased evaluation. They should cover the full range of complexity. Second, a benchmark set should reflect real-world problems. Burke et al. [48] stated there is a critical need to use real-world data more frequently for the nurse scheduling problem, to increase the implementation of algorithms in practice. Third, the size of the benchmark instances should be large enough to facilitate meaningful statistical analyses. The instance size and solvability trade-off makes a mix of smaller and larger sized instances (and hence easier and harder computational analyses) a promising option. Finally, a benchmark set should be easily extendable with other features by other researchers

[298].

Where realism, size, and extensibility are independent of the instance characteristics, a diverse benchmark set needs the identification of instance characteristics. Different combinations of instance characteristics, such as the variation of the surgery duration, result in different complexities and different solutions. Therefore, instance characteristics need to be identified, so as to be able to generate and classify diverse instances. This characterization should be generic, and applicable to any instance [299].

Any study to all instance characteristics will be very time-consuming. An alternative is to generate a benchmark set that includes various instance types within a specific area [163]. The literature describes a limited number of instance generators for systematically generating instances, such as (extended versions of) ProGen [91, 162], DaGen [1], and RanGen [84]. All these examples are of instance generators for project scheduling.

In conclusion, numerous benchmark sets exist for combinatorial optimization problems. However, for healthcare scheduling problems in general, and for the surgery scheduling problem in particular, no benchmark sets have been developed. An effective benchmark set should satisfy four conditions: diversity, realism, size, and extensibility.

8.3 Classification of surgery scheduling instances

Before generating benchmark instances, we first introduce the characteristics of surgery scheduling instances and the underlying case mix (Section 8.3.1) and then introduce a classification method for surgery scheduling instances (Section 8.3.2). Finally, we give some examples of how to incorporate specific surgery scheduling problem settings in the instances, using various instance configurations (Section 8.3.3).

8.3.1 Case mix and surgery scheduling instance characteristics

In this subsection, we give a formal description of the characteristics of surgery scheduling instances and the case mix they are based on.

We aim to include only generic instance parameters in our benchmark set, to allow other researchers to easily extend the set to include problem-specific parameters. We refer to the surgery scheduling problem classification of Cardoen et al. [52] for an extensive overview of specific surgery characteristics that can be incorporated as additional parameters of scheduling instances.

Underlying a surgery scheduling instance is the hospital's case mix, which describes the volume and properties of all surgery types that the hospital performs. A surgery type has a duration distribution, and is performed by (surgeons from a) surgical specialty. The 3-parameter log-normal distribution is proven to have the best fit with surgery duration distributions [199, 285]. Therefore, for

each surgery type $t \in T$, we use three parameters $\mu_t \in (0, \infty)$, $\sigma_t \in (0, \infty)$, and $\gamma_t \in (0, \infty)$ corresponding with the mu, sigma, and threshold (location) of the 3-parameter log-normal distribution. Using these parameters, an average duration m_t and standard deviation s_t can be calculated for each surgery type $t \in T$, using the following formulas:

$$m_t = \gamma_t + e^{\mu_t + \frac{\sigma^2}{2}} \tag{8.1}$$

$$s_t = \sqrt{\left(e^{\sigma_t^2} - 1\right) e^{2\mu_t + \sigma_t^2}} \tag{8.2}$$

In addition to a duration distribution, a surgery type $t \in T$ has a normalized relative frequency $f_t \in [0, 1]$ in the case mix ($\sum_t f_t = 1$). So, if $f_t = 0.01$, then on average 1% of all surgeries is of type t.

With a given case mix, one can generate instances of any desirable size. To generate a surgery scheduling instance requires a number of operating room blocks of given duration (e.g., 15 operating rooms of 8 hours), and a load parameter $\alpha \in [0, \infty)$ that determines the total expected surgery workload. For example a load of $\alpha = 0.9$ means that the total expected surgery duration equals 90% of the total operating room block durations. An instance with 15 operating room blocks of 8 hours and $\alpha = 0.9$ implies a total expected surgery workload of $0.9 \cdot 15 \cdot 8 = 108$ hours. It is now straightforward how to generate an instance with a total expected workload of 108 hours and given case mix characteristics. Reversely, given a set of surgeries, we can calculate the total expected surgery workload by adding all expected surgery durations.

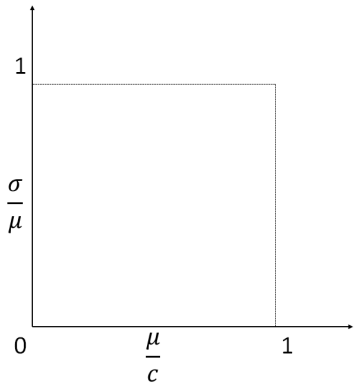
In conclusion, a surgery scheduling instance is based on a case mix, which is a collection of surgery types with a volume and duration distribution. A surgery scheduling instance consists of a list of surgeries with a corresponding surgery type. Each surgery type has a given frequency and duration distribution. An instance also contains a number of operating room blocks of given equal capacity and target load, to which surgeries should be assigned.

8.3.2 Classification of surgery scheduling instances

Section 8.3.1 explained that an instance originates from a case mix. Given a particular case mix, unlimited instances can easily be generated randomly. As each operating room department has its own case mix, a surgery scheduling benchmark set is only complete, if it encompasses the diversity of prevalent case mixes. This raises the need to classify case mixes. We propose a classification based on the surgery type duration and the coefficient of variation. Both parameters are indicators of the complexity of a scheduling problem.

The surgery type duration divided by the operating room block capacity is an indicator of the scheduling flexibility. Instances where most surgeries have a high duration, lead to schedules with gaps, since there are not enough short duration

Figure 8.1 Case mix classification visualization



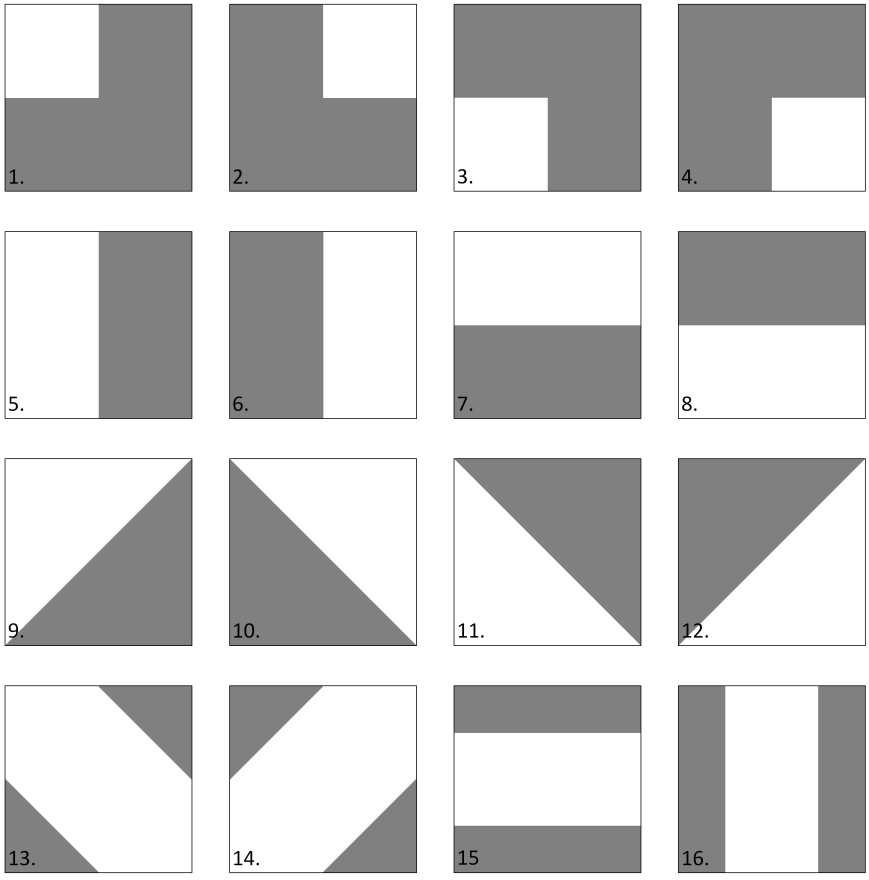
surgeries to fill these gaps. In our experience, a thorax surgery unit is such an example. Here, surgery durations are typically 4-6 hours within eight hour working days. Contrarily, outpatient surgery departments typically have low surgery durations with a high repetition, thus allowing dense operating room-schedules to be made.

The coefficient of variation is an indicator of the variability of a system and equals the standard deviation divided by the mean [293]. A higher coefficient of variation indicates high variability in surgery duration, which leads to more uncertainty in the realization of instances. Therefore, it affects the performance of realized schedules, for example in terms of overtime, utilization, or cancellations. This effect may necessitate applying a more robust approach, in which (for example) scheduled slack time is used to alleviate the effects of uncertain surgery durations [128]. A low coefficient of variation results in easier scheduling problems with almost no risk of incurring overtime and a high probability of fully utilizing operating room blocks, for example in an outpatient operating room department [293].

To classify case mixes based on these parameters, we propose the visualization shown in Figure 8.1. The x-axis is the expected duration (m) of the surgery type in relation to the total capacity of one operating room block (c). The y-axis corresponds to the coefficient of variation ($\frac{s}{m}$) in surgery type durations. Note that besides case mixes, instances can also be plotted in the case mix visualization.

In addition to the case mix classification we define so-called case mix profiles, which are partitions of the case mix classification. We consider 17 case mix profiles, each of which is such a partition. Figure 8.2 shows 16 of these case mix profiles; case mix profile 0 is the one in which all surgery types are included.

Figure 8.2 Theoretical case mix profiles – The white area indicates what surgery types are included in the case mix, following the case mix classification in Figure 8.1. The X- and Y-axes range between 0 and 1. Note: case mix profile 0 is omitted, for it contains all surgery types.



8.3.3 Surgery scheduling problem settings

Specific surgery scheduling problem settings can be introduced using various instance configurations. We discuss three examples below.

First, patient characteristics, such as urgency, can be taken into account [53]. For example, emergency patients may be classified as urgent patients. They typically have a higher variation in surgery durations. Therefore, the case mix underlying the instances representing emergency departments, could be case mix profile 1 or 2, or the combination of both, case mix profile 7.

Second, both block scheduling and open scheduling systems can be analyzed [86]. With block scheduling a range of medical disciplines can be modelled using

a combination of instances with varying case mixes. Medical disciplines can be modeled using a specific case mix per discipline. For open scheduling, the combination of several disciplines can be modeled by combining the instances following from each medical discipline. When block scheduling is considered, each medical discipline has its own instance with a set number of operating rooms, which can be optimized independently of other disciplines. The same approach can be applied to larger planning units, such as hospitals. Instances with different underlying case mixes can be combined to derive a specific combination of surgeries for a given number of operating rooms.

Finally, researchers can choose the amount of uncertainty incorporation [52]. Researchers may choose to only consider the deterministic realizations, or use the 3-parameter log-normal distribution underlying each surgery.

8.4 Example application of case mix classification

In this section, several case studies from both the literature as well as from practice are identified and classified based on the instance classification proposed in Section 8.3.

Marcon et al. [190] simulated the operating theater in order to master the risk of no realization and to maximize the operating rooms' utilization time. To analyze the performance of their approach, they generated instances consisting of surgeries with a mean case duration (contained between 60 and 180 minutes in multiples of 10), and a standard deviation (between 10% and 50% of the case duration). Their operating room opening hours were 8 hours per operating room, which clearly positions the case mix of this work in the lower-left quadrant (case mix profile 3), as shown in Figure 8.3. This case mix gives the opportunity to derive high performances on different performance indicators, for example in terms of utilization, compared to case mixes with higher coefficients of variations or longer surgery durations [293]. Lamiri et al. [170] developed methods for operating room planning with shared capacity. Their instances were based on surgeries with a uniform distributed duration on the interval [0.5, 3.0]. Since no standard deviation was included, the case mix can only be plotted at the horizontal axis, as shown in Figure 8.4. The operating room opening hours were 8 hours a day, which positions this case mix in the lower left quadrant as well.

Many studies exist where authors based test instances on real-life data from partnering hospitals. Marques et al. [193] developed a model to maximize the utilization of operating room theaters. Their instances were based on real-life data from a hospital in Portugal, as shown in Figure 8.5. Hans et al. [128] published a robust surgery loading approach for surgery scheduling using instances based on ten years of data of all elective inpatient surgeries in Erasmus MC, a large academic hospital in the Netherlands. Using this dataset, not only the hospital's case mix can be determined, but also the case mix per specialty, as shown in Figures 8.6, 8.7, and 8.8. These figures show that Erasmus MC has many

Figure 8.3 Case mix plot [190]

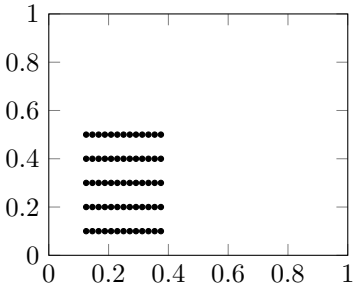
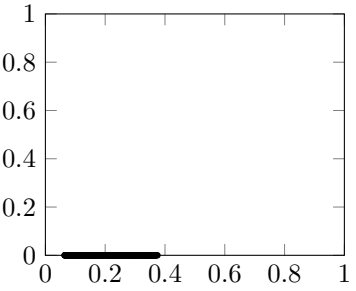


Figure 8.4 Case mix plot [170]



small surgeries compared to their opening hours, but that there are some large differences between specialties. Riise and Burke [260] used real-life data from a Norwegian hospital in order to generate test cases for their method for surgery admission planning [189], as shown in Figure 8.9. Even though the dataset contained many realizations, the surgery types were not as specific as in the Erasmus MC dataset.

Figure 8.5 Case mix plot [193]

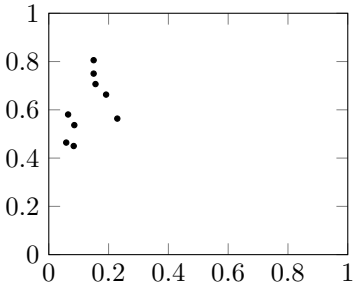


Figure 8.6 Case mix plot Erasmus MC overall

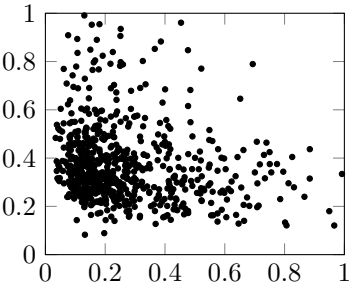


Figure 8.7 Case mix plot Erasmus MC specialty 1

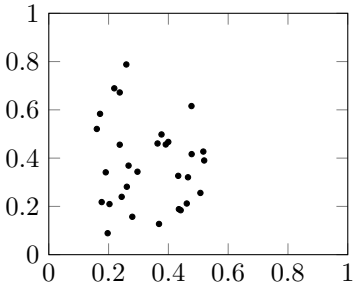
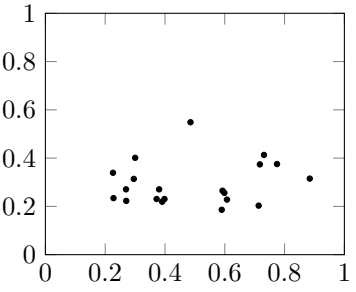


Figure 8.8 Case mix plot Erasmus MC specialty 2



Furthermore, we analyzed some datasets from both academic and non-academic hospitals in the Netherlands. The case mix of a specialized hospital is

shown in Figure 8.10. This hospital is dedicated to orthopedic surgeries, which corresponds with case mix profile 3. Figure 8.11 shows the case mix of a general hospital, and the case mix of a university hospital is shown in Figure 8.12. As expected by the more complex nature of surgeries performed in an academic hospital, their case mix has a higher coefficient of variation compared to the general hospital case mix.

Figure 8.9 Case mix plot Norwegian hospital [189]

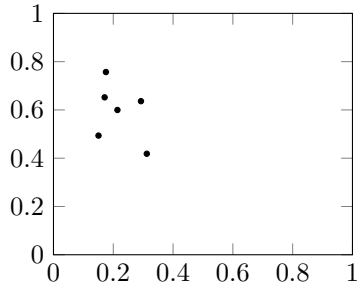


Figure 8.10 Case mix plot specialized hospital

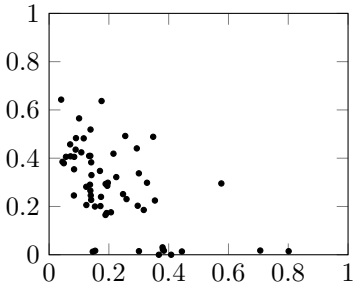


Figure 8.11 Case mix plot general hospital

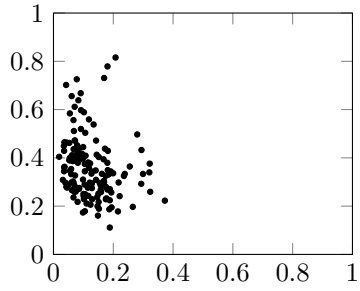
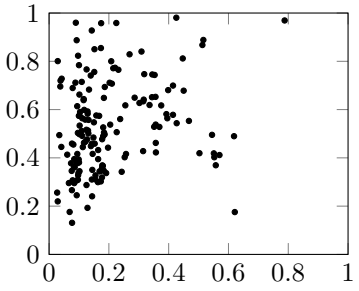


Figure 8.12 Case mix plot university hospital



As one can see, the performance of surgery scheduling algorithms for case mixes such as the first specialty of Erasmus MC, or a specialized hospital, can be analyzed using for example the generated data of Marcon et al. [190]. However, for an academic hospital, such as the Norwegian hospital, the data of Marques et al. [193] is more appropriate. Therefore, the case mix of an algorithm's instance should be classified in order to apply the results in generic or real-life settings.

8.5 Benchmark set for surgery scheduling

A benchmark set should contain a diverse selection of instances [298]. In Section 8.3 we showed how diverse instances can be identified based on their underlying case mix. This allows for generating diverse instances. This section presents the parameter settings and describes the generation of benchmark instances.

Based on a case mix and the surgery types' characteristics we can generate surgery scheduling instances, by drawing a number of corresponding surgeries. Multiple surgery scheduling instances can be combined to form an instance where surgeries of different specialties are planned in shared operating room blocks (e.g., the open block scheduling strategy), or a surgery scheduling instance forms an instance for a surgical specialty that schedules surgeries in its own operating room blocks (e.g., the closed block scheduling strategy).

To provide a benchmark set that is applicable to a broad range of surgery scheduling problems, all blocks generated have equal capacity. When an instance is needed with various block capacities, multiple instances can be combined, as explained in Section 8.3.3. All time-related data, such as surgery durations, are scaled to [0,1], in which 1 represents the default capacity of one block.

We generate two datasets of underlying surgery types. The first (Section 8.5.1) is based on a collection of real-life data from different hospitals throughout the Netherlands. The second (Section 8.5.2) is based on theoretical (generated) data. We will use these two datasets to generate benchmark sets, which consist of instances, each of which is a set of surgeries. Before presenting the final benchmark sets in Section 8.5.4, we present our instance generation and selection approach in Section 8.5.3. Here, we will present a novel technique to generate and select instances in such a way, that the resulting instances in the benchmark set are sufficiently diverse.

8.5.1 Real-life instance generation

Table 8.1 shows the experiment design for the instance generation of the real-life benchmark set. We consider the eleven specialties shown in Table 8.1, which are the most common in practice in our experience. The surgery types underlying the dataset, are derived from data from multiple academic and non-academic hospitals throughout the Netherlands over the past 10 years. To consider uniformed surgeries, we chose to include the surgery type ID, the specialty, the frequency of this surgery per year, and the μ, σ, γ, m, and s of each surgery type. Note that the relative frequency per specialty per year is considered. For ease of interpretation, in the benchmark instances, the unit of the surgery type duration distributions is minutes. Without loss of generality, we set the operating room capacity (c) to the common value of 480 minutes.

To determine the surgery types based on historical data of performed surgeries, we need to cluster individual surgeries into logistically similar surgery types. The lowest level clusters are clusters based on surgical procedure codes that correspond to patient types. As coding systems differ between countries, a higher level clustering based on logistical characteristics can be applied to allow the comparison of case mixes (e.g., surgeon or surgical specialty, surgery duration).

We have derived over 1,000 surgery types from almost 200,000 surgery realizations from 5 hospitals, from recent years. For each anonymized surgery in this set, the surgical specialty, treatment code, and duration realization was provided. We cluster these surgery realizations based on their specialty and treatment code, a

commonly used classification in the Netherlands to indicate planning characteristics. To obtain surgery types, we fit a 3-parameter log-normal distribution to each cluster of surgery realizations, using a mean squared error (MSE) minimization procedure. Surgery types were only included when more than 20 realizations are available, and when the derived distribution's MSE is smaller than 0.001. Table 8.1 shows the statistics of the outcome.

Table 8.1 Surgery realizations underlying real-life specialties

Specialty short name	Specialty full name	No. surgery types	No. surgery realizations
CHI	General surgery	149	9,311
ENT	Otolaryngology	146	11,986
EYE	Ophthalmic surgery	91	7,953
GYN	Obstetric and gynecologic surgery	60	4,116
MIX	Remaining specialties, including colorectal surgery, pediatric surgery, trauma surgery, vascular surgery, etc.	173	46,938
NEU	Neurological surgery	47	2,832
ONC	Surgical oncology	43	6,466
ORT	Orthopedic surgery	133	7,618
PLA	Plastic surgery	73	3,022
THO	Thoracic surgery	28	2,248
URO	Urology	75	5,627
Total		**1,018**	**108,117**

We consider eight values for the number of operating rooms ($j \in \{5, 10, 15, 20, 25, 30, 35, 40\}$). Surgery scheduling in practice typically has a planning horizon anywhere between a single day and two weeks. In almost all hospitals we collaborate with, the planning horizon is one week, during which specialties typically have up to 20 blocks of surgery time.

We consider ten values for the ratio of surgery load to operating rooms ($\alpha \in \{0.80, 0.85, \ldots, 1.20\}$). An instance is considered of a certain surgery load if the deviation of the surgery load is less than 0.025. For example, an instance with surgery load $\alpha = 0.80$, should have a workload between 0.775 and 0.825.

For each specialty, we provide a total of X instances. The instances are built by repeatedly selecting a random surgery type, for which we add one patient (surgery) to a set of patients until the ratio of surgery load to operating rooms deviates less than 0.025 from the desired α. While the ratio is less than α, we do 100 attempts to add another random patient, who is added iff the resulting ratio is closer to the desired α. Accordingly, we generate $3X$ instances, from which we finally select X instances for the benchmark set. The instance generation procedure is summarized in Box 1. In Section 8.5.3 we explain how we perform this selection, aiming to select the X most diverse instances.

Table 8.2 Real-life instance generation design [85]

Problem parameter	Values considered	Number of values
Specialty (p)	CHI, ORT, ENT, GYN, PLA, URO, EYE, THO, ONC, NEU, MIX	$\|P\|$=11
Number of operating rooms (j)	5, 10, 15, 20, 25, 30, 35, 40	$\|J\|$=8
Load (α)	0.80, 0.85, ..., 1.20	$\|A\|$=10

Total instance parameters: 880

With $X = 10$ instances for all combinations of parameters, the final real-life benchmark set consists of 8,800 instances, as shown in Table 8.2.

8.5.2 Theoretical instance generation

In this section we describe how we generated the benchmark sets with theoretical instances. Instead of generating instances for specialties (for the real-life instances), here we focus on the case mix profiles described in Section 8.3.2.

Table 8.4 shows the experiment design for the generation of theoretical instances, which leads to 1,360 instance parameter combinations. In comparison to the real-life instances, observe that the only difference is that here we use 17 case mix profiles, instead of 11 specialty case mixes. The instance generation procedure is almost the same as in the previous section. The difference lies in how we select a random surgery type. For the theoretical benchmark instances we generate a random surgery type for every surgery we (try to) add to an instance. The procedure is as follows. For each surgery t we sample a random coordinate (X_t, Y_t) from the case mix profile at hand. Consequently, X_t is the expected duration in relation to the operating room capacity ($\frac{m_t}{c}$), and Y_t is the coefficient of variation ($\frac{s_t}{m_t}$). So, we have that $m_t = cX_t$ and $s_t = X_tY_t$. Now, from m_t and s_t we must determine the three parameters $\mu_t \in (0, \infty)$, $\sigma_t \in (0, \infty)$, and $\gamma_t \in [0, \infty)$, corresponding with the mu, sigma, and threshold (location) of the 3-parameter log-normal distribution respectively. Equations ((8.1)) and ((8.2)) describe the relation between m and s and μ_t, σ_t, and γ_t. Since we have one unknown parameter too many to solve the equations, we choose to randomly set $\gamma_t = 0.75Rm$, where R is a random number in [0,1]. This formula follows from our analysis of the threshold parameters of the real-life surgery types used in the previous section, which we found to lie between 0 and 0.75.

The instance generation is equal to the procedure of the real-life instance generation, as summarized in Box 1. Again, $3X$ instances per combination of parameters are generated, from which the X most diverse instances are selected, as explained in Section 8.5.3. With $X = 10$ instances per parameter combination, the theoretical benchmark set consists of 13,600 instances. In total, the real-life and theoretical instances in the benchmark set amount to 22,400 instances.

Table 8.4 Theoretical instance generation design [85]

Problem parameter	Values considered	Number of values
Case mix profile (p)	0,1, ..., 16	$\|P\|$=17
Number of operating rooms (j)	5, 10, 15, 20, 25, 30, 35, 40	$\|J\|$=8
Load (α)	0.80, 0.85, ..., 1.20	$\|A\|$=10

Total instance parameters: 1,360

Box 1 *Instance generation procedure*
Given:

- a case mix,
- number of operating rooms,
- a set of loads ($\alpha \in \{0.80, 0.85, \ldots, 1.20\}$),

the following instance generation procedure will generate $3X$ instances for each of the loads in the set:

1. Select a random surgery type
2. Add a surgery of this type to the instance, and determine the load (total expected surgery duration divided by the total operating room capacity) of the instance.
3. *If* the load deviates less than 0.025 from any load α in the set, continue with Step 4.

   Else, *if* the load is more than the highest load α+0.025, discard the instance and continue with Step 6. *Else* repeat Step 1-3.
4. If the load is less than the desired load α, do 100 attempts to add another random surgery, which is added iff the resulting load is closer to α.
5. Save the resulting instance for the desired α. If a load α has sufficient instances, remove it from the set of loads.
6. Repeat the generation procedure until for each load α in the set, $3X$ instances are generated.

8.5.3 Instance proximity maximization

Our aim is to generate a benchmark set with instances that are mutually significantly different. A benchmark set serves to give insight into the problem characteristics that make it hard for an algorithm. An instance that is very similar to another instance in the set thus provides no additional insights. To

the best of our knowledge there exists no measure for assessing the similarity of instances, besides data mining techniques. We therefore propose the following approach to measure similarity of instances, which we shall refer to as instance proximity.

Consider two instances A and B, which were generated based on the same case mix, which have the same load and the same number of operating rooms. As an instance consists of a set of surgeries, we need to assess the similarity of surgeries in instance A and B. Surgeries are characterized by 5 aspects: the expected duration, duration standard deviation, three parameters of the log-normal duration distribution (see Section 8.3.1). To assess the similarity of two surgeries say S1 and S2 from respectively instance A and B we could either take a deterministic or stochastic approach. A deterministic approach would only consider the expected duration. One could for example say that if the absolute difference between the expected durations of S1 and S2 is below a threshold, the surgeries are regarded proximal. In a stochastic approach we would consider the duration distribution characteristics of S1 and S2. In this case, two surgeries are regarded proximal if the overlap of the duration distributions' density functions is above a threshold. Observe that the deterministic approach discriminates less than the stochastic approach. For example, two surgeries may be considered proximal from a deterministic point of view, while not being considered proximal from a stochastic point of view (i.e., their expected durations are proximal, but distribution functions overlap insufficiently). Since using a deterministic approach will yield less proximate instances, in the remainder we shall apply the deterministic approach. We introduce the following definition:

Definition 1. *Surgeries S1 and S2 are ϵ-proximate iff their expected durations differ less than ϵ%.*

In order to select those instances that are maximally different, we first analyze the ϵ-proximity of all surgery pairs between two instances. Observe that surgeries can be ϵ-proximate to more than one other surgery from the other instance. Therefore, to determine how proximal two instances are, we have to find the maximum matching of ϵ-proximate surgeries, which can be found in polynomial time. The so-called ϵ-proximity quantity of two instances is determined as follows:

Definition 2. *The ϵ-proximity quantity of two instances equals the total workload of all ϵ-proximate surgeries selected in the maximum matching from both instances, divided by the total workload of all surgeries from both instances.*

We evaluate the ϵ-proximity quantity ($a_{i,j}$) for every combination of instances (i and j). When all instances are generated and compared, we select the X instances among which the maximum proximity (Z) is minimal. We do this by solving the following ILP. Alternatively, the Ford-Fulkerson algorithm or Hopcroft-Karp algorithm could be used.

$$\min Z \tag{8.3}$$

$$s.t. \sum_i Y_i = X \tag{8.4}$$

$$Q_{i,j} \leq Y_i \qquad \forall i,j \tag{8.5}$$

$$Q_{i,j} \leq Y_j \qquad \forall i,j \tag{8.6}$$

$$Q_{i,j} \geq Y_i + Y_j - 1 \qquad \forall i,j \tag{8.7}$$

$$a_{i,j} Q_{i,j} \leq Z \qquad \forall i,j \tag{8.8}$$

$$Q_{i,j} \in \{0,1\}, \quad Z \geq 0, \quad Y_i \quad \in \{0,1\} \qquad \forall i,j \tag{8.9}$$

Here, binary variable Y_i indicates whether we select instance i, binary variable $Q_{i,j}$ is 1 iff both $Y_i = 1$ and $Y_j = 1$ (i.e., instances i and j are both selected). These selected X instances are thus the least ϵ-proximal within the original instances set, and therefore the X most different instances. In Appendix I we show some statistics of this procedure, and we summarize this proximity maximization procedure in Box 2.

Box 2 *Proximity maximization procedure*
Repeat the following proximity maximization procedure for every parameter combination.

1. Generate $3X$ random instances.
2. Determine for all surgery pairs of all pairs of instances if they are ϵ-proximal.
3. Determine the maximum matching of ϵ-proximate surgeries.
4. Determine the ϵ-proximity quantity of all pairs of instances.
5. Use the ILP to select X instances for which the maximum ϵ-proximity quantity between all instance pairs is minimal.

8.5.4 Benchmark set

The benchmark set satisfies all four conditions of Vanhoucke and Maenhout [298]: Diversity, realism, size, and extendibility. The instances are based on a wide range of case mix types, either based on real-life case mixes from one of 11 surgical specialties, or based on one of 17 theoretical case mix profiles. In addition, $|J| \cdot |A|$=80 variations of problem size characteristics are used, as shown in Table 8.2 and Table 8.4. As a result, the benchmark set is very *diverse*. This is further strengthened by our instance generation procedure, in which we use the concept of instance proximity to maximize instance diversity. In Appendix II we show some statistics on the case mix diversity of the generated instances.

The benchmark set reflects *real-world* problems by the use of the underlying patient type dataset. The instances differ in *size*, to facilitate a benchmark set with a mix of smaller and larger instances for solvability reasons. Other researchers can *extend* the benchmark set with their own characteristics. Also, the instance generator is provided on the website, for the reader to generate even more instances.

To enable researchers to assess a solution method's performance using a smaller set, we selected a subset from the total benchmark set of 22,400 instances. This smaller benchmark set consists of those instances that we were not able to solve to optimality within 10 minutes using CPLEX for the two variants of the surgery scheduling problem introduced in Appendix III. Furthermore, in this subset one instance per parameter combination was selected, and only instances of average loads (0.95, 1.0 and 1.05) are considered.

The benchmark set, the benchmark subset, the instance solution validator, our first solutions, and the instance generator are available for the academic community at:

https://www.utwente.nl/choir/en/research/BenchmarkORScheduling/

Instance and solution files are in the plain ASCII text format (tab separated), and detailed descriptions of their formats are provided on the website. To facilitate statistical and sensitivity analyses based on the benchmark set, a sample generator is also provided at this URL. For each surgery in each instance, durations can be sampled from the 3-parameter log-normal distribution of the surgery type's distribution. These samples can be used for simulation purposes.

Appendix III gives more details on the specifics of the surgery scheduling problem and the first solutions that we provided on the website.

All programs were developed in the Embarcadero® Delphi XE8 programming language, and compiled to MS Windows executables.

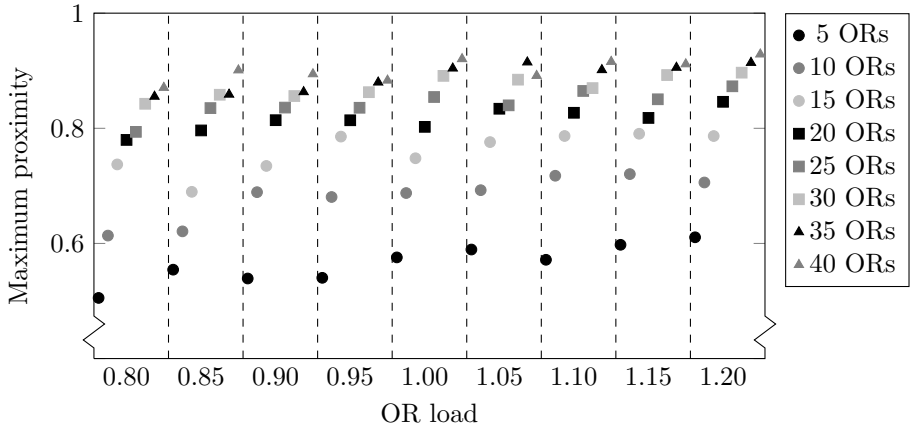
8.6 Conclusions and discussion

Benchmarking the performance of surgery scheduling between different hospitals is difficult, as case-mixes differ between hospitals. We have proposed a benchmark set and a case mix classification to facilitate (benchmarking) experiments. We have also developed a novel instance generation procedure that maximizes the difference between instances.

The proposed generic benchmark set for surgery scheduling algorithms is diverse, derived from real-world data, varies in size, and is extendable, according to the characteristics of an effective benchmark set [298]. The benchmark set, the instance generator, and solution validator can be downloaded from the website:

https://www.utwente.nl/choir/en/research/BenchmarkORScheduling/

Figure 8.13 Maximum proximity of selected instances of case mix type 2 in benchmark set



On the website we also provide a small benchmark set consisting of a subset of hard instances of the large benchmark set. We also provide our initial solutions to all instances, and a sample generator.

We found that the diversity in the real-life case mixes is much higher than for generated (theoretical) data found in the literature. Therefore, further research is needed to analyze the relation between hospital case mixes, and case mixes used in literature. Furthermore, suitable algorithms for instances with specific case mixes can be developed. For example, we already mentioned that robust approaches are more suitable for solving instances with an underlying case mix with a high coefficient of variability.

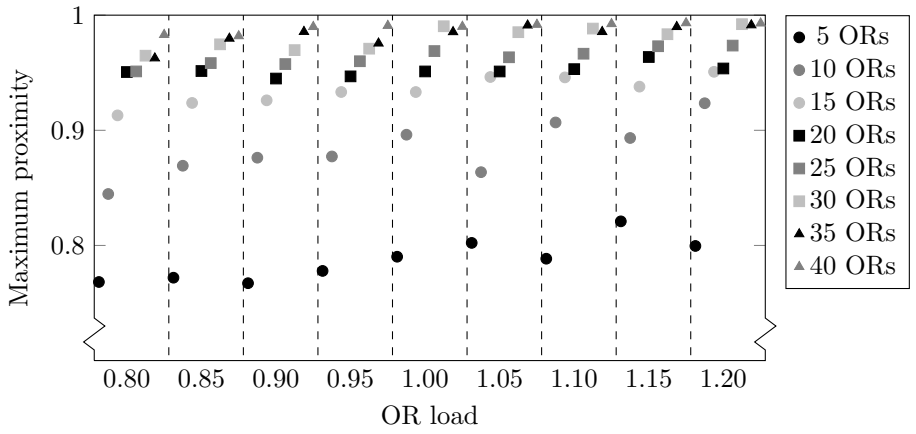
The proposed case mix classification is visual, and gives insight into what type of case mix is under consideration. Practitioners can use the case mix classification to get insight in to what extent another case mix (than their own) is comparable to their own. Hence, given for example instances or case mixes found in the literature, a practitioner can assess to what extent the results are applicable in their own situation.

8.7 Appendix I

An analysis of the proximity statistics of the instances in the benchmark set shows that the maximum proximity of instances is influenced by the case mix and the problem size (number of operating rooms). Figure 8.13 and Figure 8.14 show proximity statistics of theoretical case mix profile 2 and 3, respectively. The X-value of each dot represents the instance parameter combination. The Y-value is the maximum proximity for 10 instances that were selected from 30 random instances using the ILP-procedure in Section 8.5.3.

The figures show that the instances from theoretical case mix profile 2 are

Figure 8.14 Maximum proximity of selected instances of case mix type 3 in benchmark set



less proximal than those of theoretical case mix profile 3. This can be explained by the range of the possible distributions that can be selected in a case mix. In theoretical case mix profile 3, all distributions have a small σ and small μ, whereas in theoretical case mix profile 2, the σ and μ are both larger. Furthermore, in theoretical case mix profile 3, only small shift values can be derived, whereas in theoretical case mix profile 2 both larger and smaller shift values can occur.

As would be expected, the figures also show that a higher number of operating rooms affects the maximum proximity. More operating rooms, means more surgeries, and thus a higher possibility of drawing two similarly distributed surgeries.

The proximity statistics of all instance sets in the benchmark show similar trends. In the benchmark set we have included the detailed proximity statistics for all these instance sets (i.e., for each case mix profile and each parameter combination).

8.8 Appendix II

To show the case mix diversity of the benchmark instances, we plotted several generated instances based on specialty case mixes in the instance classification plot (see Figure 8.15 and Figure 8.16). Furthermore, we plotted several instances based on theoretical case mix profiles (see Figure 8.17 and Figure 8.18).

For the theoretical instances, one can easily detect the corresponding case mix profile underlying the instance. For example Figure 8.17 was based on case mix profile 5, with surgery types in the left half, whereas Figure 8.18 was based on case mix profile 13. The real-life instances are harder to differentiate. Figure 8.19 shows the underlying specialty case mix of the instance of Figure 8.15. Now, one can see that the instance corresponds with its underlying case mix.

Figure 8.15 Instance plot of ONC instance

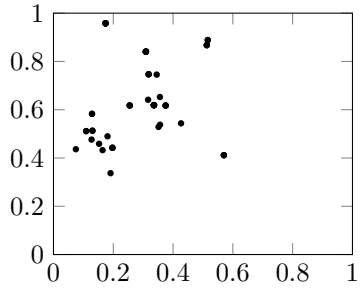


Figure 8.16 Instance plot of URO instance

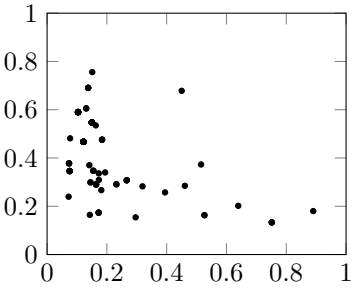


Figure 8.17 Instance plot of theoretical instance no. 5

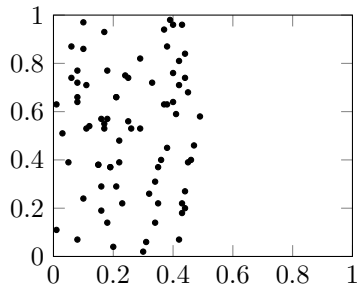


Figure 8.18 Instance plot of theoretical instance no. 13

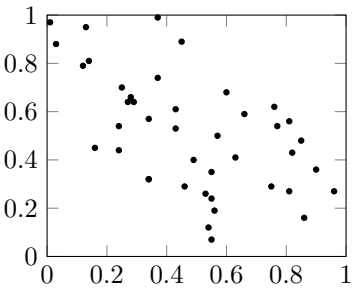
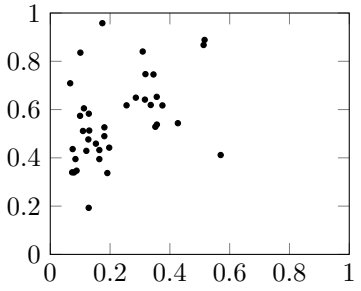


Figure 8.19 Underlying case mix of ONC instance



8.9 Appendix III

The large amount of surgery scheduling literature consists of many problem variants. To provide some first solutions to the benchmark set developed in this chapter, we present two basic surgery scheduling problem variants. Both variants consider the idle time of an operating room as performance indicator, together with a variant specific performance measure.

Variant A: Planning surgeries in overtime, also known as overbooking, is not

allowed. Therefore, the sum of the expected duration of all canceled patients is an important performance measure to this variant.

Variant B: All surgeries need to be scheduled, if necessary in overtime. Therefore, the planned overtime is an important performance measure to this variant.

8.9.1 Formal problem formulation

Before we define the problem, we first introduce some notation, as shown in Table 8.5. Let $s \in S$ be the set of surgeries, and $j \in J$ the set of operating rooms. c_j is the capacity of operating room j, and $X_{s,j}$ is a binary variable that indicates whether surgery s is scheduled in operating room j. Recall that μ_s is the expected duration of surgery s. To be able to identify an unscheduled surgery, we introduce Y_s, which is a binary variable indicating whether a surgery s is unplanned ($Y_s = 1$) or planned ($Y_s = 0$).

Table 8.5 Notation

Set, parameter or variable	Definition
$s \in S$	Set of surgeries
$j \in J$	Set of operating rooms
c_j	Capacity of operating room j
μ_s	Expected duration of surgery s
$X_{s,j}$	Binary variable indicating whether surgery s is scheduled in operating room j (1) or not (0)
Y_s	Binary variable indicating whether surgery s is unplanned (1) or planned (0)
I_j	Idle time of operating room j
C	Sum of the expected durations of all canceled surgeries
O_j	Expected overtime of operating room j

We want to minimize the idle time per operating room (I_j), and depending on the problem variant the expected duration of all canceled patients (C) or the planned overtime per operating room (O_j). Note that for variant A holds that $O_j = 0$, and for variant B holds that $C = 0$. This gives the following objective:

$$\min C + \sum_j I_j + O_j.$$

We identify the following constraints:

$$\sum_j X_{s,j} + Y_s = 1 \quad \forall s \in S,$$

which forces each surgery to be planned at an operating room, or be canceled.

We define the idle time, number of cancellations, and the overtime as follows:

$$I_j = \left[c_j - \sum_s X_{s,j}\mu_s \right]^+ \quad \forall j \in J,$$
$$C = \sum_s Y_s \mu_s,$$
$$O_j = \left[\sum_s X_{s,j}\mu_s - c_j \right]^+ \quad \forall j \in J.$$

The idle time of an operating room is the difference between the capacity of the operating room and the sum of the scheduled durations at the operating room. The expected duration of all canceled patients equals the sum of all patients that are unscheduled times their expected duration. The overtime is the difference between the sum of the scheduled durations at the operating room and the capacity of the operating room. This gives the following constraint:

$$\sum_s X_{s,j}\mu_s = c_j + O_j - I_j \quad \forall j \in J. \tag{8.10}$$

To distinguish between variants A and B, we add variant specific constraints. To ensure no patients are scheduled in overtime in variant A, we add the following constraint to the problem in variant A:

$$O_j = 0 \quad \forall j \in J. \tag{8.11}$$

An operating room cannot have patients scheduled in overtime. To ensure all patients are scheduled in variant B, we add the following constraint to the problem in variant B:

$$Y_s = 0 \quad \forall s \in S. \tag{8.12}$$

As no patients are allowed to be canceled, all patients are forced to be assigned to an operating room. Finally, we have some non-negativity constraints:

$$X_{s,j} \in \{0,1\}, Y_s \in \{0,1\}, I_j \geq 0, C \geq 0, O_j \geq 0 \quad \forall s \in S, j \in J. \tag{8.13}$$

8.9.2 Solution method

To find first solutions to the surgery scheduling problem variants A and B for the benchmark instances presented in this chapter, we apply the well-known list scheduling heuristic with multiple machine and job selection rules. This heuristic has been widely used in the literature for surgery scheduling [207]. The machine selection rules are best fit (BF), first fit (FF), random fit (RF), and worst fit (WF). The job selection rules are ascending order of expected surgery duration (Asc), descending order of expected surgery duration (Des), and random selection (Rnd). This yields 12 list scheduling variants, which we denote by `Dur_Asc_BF`, ..., `Dur_Rnd_WF`.

PART 5

conclusion

Why Wait?
Organizing Integrated Processes in Cancer Care

CHAPTER 9

The impact of Operations Management in practice

9.1 Introduction

An efficient organization of healthcare processes is extremely relevant to clinical practice. This thesis describes how several of the problems faced by healthcare managers, in particular those at UMC Utrecht, can be solved theoretically. However, healthcare operations management (OM) research should not stop after theoretical results, as for a hospital, impact is made in practice.

This chapter analyzes our research center's efforts to make an impact in practice with research activities. Since its founding in 2007, our research center CHOIR (Center for Healthcare Operations Improvement and Research) has been developing an organizational structure and working methods to share knowledge, and to promote and facilitate impact. This so-called 'CHOIR ecosystem' is a close collaboration with the healthcare sector, involving research, education, and impact activities. This chapter describes the ecosystem, which follows the researcher-in-residence model. Furthermore, we discuss our experiences – most of them are based on the collaboration with UMC Utrecht – and try to identify under which conditions researchers should perform their research, to maximize the likelihood of having true impact.

This chapter is organized as follows. First, Section 9.2 discusses how healthcare organizations can engage with process optimization. Section 9.3 focuses on the underlying methodologies of these process optimization activities from an academic point of view. Then, Section 9.4 describes the network of healthcare institutions with whom CHOIR facilitates the CHOIR ecosystem. Section 9.5 reflects on the impact made in practice with the projects related to this thesis, and analyzes the critical success factors. Finally, in Section 9.6 we draw conclusions and discuss the results.

9.2 Process optimization approaches

Within the CHOIR network, we collaborate with our healthcare partners on process optimization and improvement projects. But how can hospital adminis-

trators get started with these projects? In this section we identify two complementary strategies that are used for process improvement in practice: bottom-up (Section 9.2.1) and top-down approaches (Section 9.2.2). In Section 9.2.3 we argue that a joint approach, in which the complementary bottom-up and top-down approaches are combined, provides the best results.

9.2.1 Bottom-up approaches

Bottom-up process optimization considers process improvement projects that are initiated by front-line staff, and that focus on iterative process improvements on the operational level of control, to promote a culture of continuous quality improvement.

There are various well-known improvement concepts, such as Lean Management [323], the PDCA quality-cycle [288], Six Sigma [246], Theory of Constraints (TOC) [112], Total Quality Management (TQM) [80], and Value Based Healthcare [242]. While most of these concepts or their principles originate from industry, they have increasingly been introduced in healthcare organizations in the past decade, driven by the necessity of healthcare providers to operate more efficiently and effectively. The presentation of each of the concepts differs, as most of the concepts come with their own set of tools, methodology, and techniques to visualize processes and performance. However, there is much overlap in the concepts' goals [316]. We can distinguish three common OM principles on the basis of the improvement concepts:

1. To maximize value while minimizing waste Adding value gives an organization a right to exist, and efficiency is needed for the organization to be sustainable. The time, money, and effort that is put in non-value adding activities or in inefficient activities, could be better used to provide better care for more patients. Therefore, increasing value adding activities and optimizing efficiency is important for organizations. Well-known concepts that focus on this principle are Lean Management and VBH.

2. To reduce variability Variability results in inefficiencies and reduced quality of service and labor. It results, for example, in fluctuating workloads, access time and waiting time for patients. Therefore, organizations should strive for a constant and predictable workload, close adherence to protocols, minimal errors or rework, and flexible resources. Six Sigma is an example of a concept that focuses on variability reduction. Furthermore, the design of clinical pathways reduces variability, which focuses on designing standardized protocols for the patients' clinical course. Planning, scheduling, and forecasting have a big impact on variability as well, in particular variability in workload.

3. To reduce complexity Complexity of processes results in poor management control, thus leading to inefficiency and ineffectiveness. A highly complex planning may potentially be better, and may reduce variability, but

is extremely hard and costly to implement, and might be sensitive to changes in the system. A concept that focuses on complexity reduction is for example Total Access, which instructs to do today's work today by organizing processes flexibly without the need to plan activities, for example by organizing clinics on a walk-in basis.

Given that OM in healthcare is still growing, there are a lot of process improvement opportunities. Therefore, it is important that healthcare organizations create an atmosphere of continuous quality improvement, in which all staff on all levels of control is involved in continuous improvement efforts. As an example, the concept of Lean Management strives to create such an atmosphere, and there is much evidence of its effectiveness [314]. Lean Management embodies aspects of all three of the aforementioned OM principles, as it aims to maximize value, reduce all kinds of wastes, and standardizes processes.

There are a few drawbacks to bottom-up process improvement projects. Although bottom-up approaches engage front-line staff in continuous improvement initiatives, it is hard to change planning and control on higher managerial levels. Particularly tactical and strategic planning are therefore underexposed. As these planning levels heavily impact the operational level, the impact of operational improvement concepts is constrained by the choices made at a tactical planning level. As an example we mention the block schedules for outpatient clinics and operating rooms, which determine on which day, which specialty or specialist is working. Although operational initiatives might be able to slightly decrease the workload variability, these block schedules cause great variability in workload in the integral care processes, which are repeated every week of the year, with very little changes over the years [145].

The second drawback of bottom-up approaches is that there is limited scientific evidence of their performance, besides case studies. Therefore, the motivation for the choice for a specific improvement concept by healthcare organizations is often based on a successful application in another healthcare institute [134, 316]. As it is hard for healthcare decision makers to assess which improvement concept to apply in their organization, benchmarking initiatives can contribute to stimulate healthcare providers to learn from better or best practices [224]. As healthcare practitioners are educated to practice evidence-based medicine, benchmarking effectively provides evidence of other (better) working procedures. However, one should keep in mind that benchmarking itself has some drawbacks as well. For example, improvement opportunities can still be present for the best performing institute. Furthermore, copying peer behavior makes healthcare providers become similar, which may be undesirable in a competitive healthcare sector, and may suppress innovation.

A third drawback of bottom-up approaches is the danger of over-standardization. There are many care pathways that cannot be (entirely) standardized, such as for patients with rare tumors in oncology. Processes that can be standardized, can be planned centrally, and are most suitable for efficiency gains. However, processes that cannot be standardized should allow clinicians sufficient

planning autonomy to flexibly act upon any emerging situation. Although clinical pathways make no implications for planning and control, the implementation of a clinical pathway is often done in conjunction with reserving capacity slots or blocks for groups of patients belonging to a clinical pathway, for example to promote one-stop-shopping. Observing a trend of increasing co-morbidity and more individualized care pathways (see Chapter 10), planning and control concepts, for example based on interchangeable modules, are needed to cope with patients with high care pathway variability. Block reservation of capacity for regular patients reduce the required flexibility for this new group of patients.

9.2.2 Top-down approaches

Top-down process optimization considers process improvement projects initiated by the management of an organization or department, and is often executed by a project group. Top-down approaches revolve around the re-design and optimization of processes, and planning and control. In terms of the three OM principles of Section 9.2.1, planning and control is designed to achieve an organization's goals regarding value-adding and efficiency. Planning and control can reduce variability, however it often increases complexity at the same time. Therefore, organizations need to make a trade-off between these two aspects.

Top-down approaches span decisions on all hierarchical layers in an organization. The control of the processes involved can be categorized by the framework for healthcare planning and control of Hans et al. [129]. In this thesis, we focus on the resource capacity planning, which considers the planning and control of all non-renewable resources in a healthcare institution (e.g., staff, rooms, and equipment).

Strategic resource capacity planning focuses on long-term decision making, based on long-term demand forecasts. It has the highest level of capacity and planning flexibility, but also the highest level of uncertainty due to the long planning horizon. In the Netherlands, strategic planning, especially in the form of case mix planning, is often financially driven, based on negotiations with insurers. Strategic planning addresses furthermore the dimensioning of capacity, workforce planning, and staff training.

Tactical resource capacity planning considers the organization of care in the intermediate planning horizon, which is typically several weeks or months, depending on the capacity flexibility and demand characteristics. Tactical planning has a planning horizon that, contrary to operational planning, allows capacity flexibility. Examples are temporary capacity expansions (e.g., through extra shifts, planned overtime, or temporary reallocation of staff), and changes in block schedules and department agendas. This flexibility allows for dealing with fluctuations in demand. At this level, demand is partly known, and may partly need forecasting. Tactical planning addresses questions such as template design, block allocation, multi-department scheduling, staffing, resource pooling, and admission planning. This level is the most overlooked area of planning and control in healthcare.

Offline operational resource capacity planning considers proactive decision making within a short planning horizon. Here, elective demand is known, and capacity has little or no flexibility. Offline operational decisions are for example surgery and appointment scheduling.

Online operational resource capacity planning concerns the monitoring and control of real-time processes. At this level of control, all uncertainty is materialized. Decisions at this level are made ad hoc, such as the coordination of emergency requests, the usage of the slack OR time, and the handling of patient requests when they arrive.

Top-down process improvement projects enable organizations to make structural changes to their organization, and to invest in new solutions, especially on the strategic and tactical level of planning and control. These projects are often well organized and (financially) supported by top-management as the expected impact is large.

There are various drawbacks of top-down approaches. First, through the temporary nature of most projects, it is hard for top-down interventions to create a sustainable impact that lasts on the long term, especially if it involves changing the behavior of people.

Second, in line with the first drawback, it is hard to engage staff in the organizational changes. Staff members are often needed for the implementation of the interventions. Note that the structure of a healthcare institute differs from industry, as medical doctors are autonomous individuals. Therefore, in top-down improvement projects, it is important to have the medical staff involved at the managerial level in the project, as they are important decision makers besides the hospital's management. Furthermore, the front-line staff needs to be included, as otherwise support among end-users can be low, and implementation of project results might be compromised. As front-line staff is forced to cohere to a decision made by higher level administrators, they might feel as if their interests and practical experience are neglected. As an example, consider the introduction of a nursing flex-pool, where a pool of nurses is flexibly assigned to the wards that are in need of extra personnel. However, this requires nurses to now develop a flexible skill set applicable to multiple wards, whereas they might have deliberately chosen to be trained as a specialized nurse for a specific nursing ward.

A third drawback is the complexity of top-down interventions. These interventions have a large impact on hospital-wide processes, and their consequences are hard to assess prospectively, as they are widespread in many areas. For example a tactical level intervention such as a redesigned block schedule of operating rooms has a great impact on up- and downstream processes: it necessitates redesigning the outpatient clinics' agendas, and also impacts the staffing levels of nursing wards.

9.2.3 Joint optimization

As Sections 9.2.1 and 9.2.2 showed, bottom-up and top-down approaches are opposite approaches with their own advantages and disadvantages. Specific people may favor one approach over the other, depending on expertise and interest. Top-down approaches are for example favored by OM/OR scientists and administrators, whereas bottom-up approaches are often favored by consultants and staff in the primary process. However, bottom-up and top-down approaches are in fact complementary, and strengthen each other.

Bottom-up approaches aim to create a culture of continuous improvement, which creates a focus on operational processes and performance by front-line staff. Top-down approaches on the other hand can result in higher level organizational changes. However, when operational level improvement opportunities are still abundant, front-line staff support for a top-down approach will be low. Therefore, the operational processes should operate sufficiently before initiating a top-down approach. As an example we mention an operating theater, where the Master Surgery Schedule (MSS) needs to be optimized to reduce peaks in downstream ward occupation, which is a tactical level issue, while a high number of operational problems are perceived, such as overtime, cancellations, and no-shows. Operating room staff will then consider a change in the block schedule to be invasive in their agendas, while their operational problems are not even addressed by this intervention. As a result, their support might be low.

Top-down approaches concerns the (re-)design and optimization of planning and control, in order to reach certain target operational performance levels. The first step is thus to establish this target operational performance. In our experience, key performance indicators (KPIs) are often lacking or incomplete, and hospitals do not –or at best to a limited extent– regularly measure their operational performance. Bottom-up approaches are needed to strengthen top-down approaches, as their focus is on defining, structurally measuring, and continuously improving operational performance. Moreover, bottom-up approaches engage front-line staff in operational improvement, which is a step-up to system redesign and optimization.

Bottom-up and top-down approaches are therefore required to go hand-in-hand. This also includes managerial support for continuous improvement initiatives, and front-line staff to join high-level decision-making sessions. This way, bottom-up approaches can be fruitful within the boundaries of the top-down planning and control.

9.3 Methodology

This section provides a more detailed description of the underlying scientific methodologies that we use in top-down as well as a bottom-up process optimization research activities. We follow the steps of the Managerial Problem-Solving Method (MPSM), which is the engineering approach for business problems.

9.3.1 Managerial Problem-Solving Method

The MPSM is a systematic and efficient methodology to identify and efficiently deal with problems, using an engineering approach. Within a step-by-step approach, the MPSM allows for creativity and flexibility, which is required for engineering solutions to reach high quality solutions. It consists of seven phases [137]:

1. Defining the problem;
2. Formulating the approach;
3. Analyzing the problem;
4. Formulating (alternative) solutions;
5. Choosing a solution;
6. Implementing the solution;
7. Evaluating the solution.

In each of the seven phases, multiple tools and techniques can be applied, depending on the researcher's knowledge and skills set. Furthermore, more or less attention to each of the seven phases can be given, depending on the need in practice. This makes the MPSM suitable to use as a framework for healthcare practitioners and students in all stages of their studies.

The remainder of this section addresses several aspects of the MPSM phases in more detail. Phase 1 of the MPSM aims to find the root causes of the original perceived problem, by constructing a problem bundle. This enables researchers to extract the research problem by demarcating the research scope to a subset of the root causes, as explained in Section 9.3.2. In Phase 1 also the concept performance is translated into measurable performance indicators. By comparing the initial performance with the (prospectively assessed) performance after the interventions in Phase 5, the performance improvement can be objectified. We elaborate on this in Section 9.3.3. Phases 4-7 revolve around designing, testing and implementing solutions. Section 9.3.4 elaborates further on this.

9.3.2 Problem bundle and demarcation of core problem

Problems of healthcare organizations that they want to be solved are often perceived problems, and the problem at hand might just be a consequence of another problem. Therefore, when starting on an improvement project, we must first collect, analyze and quantify all problems related to the perceived problem, until finding the core problem that has no direct cause itself, and that can be influenced. This requires intensive collaboration with the problem owners, as well as a data analysis. A problem cluster helps to structure the context of the problem and the causal relations between the problems, and helps to select the core problem.

In many cases there are multiple core problems. A high rejection rate at the wards might for example be caused by an unbalanced OR schedule, as well as unbalanced nurse staffing. As a researcher, we must choose a core problem that can be influenced. For example, when we collaborate with people from the wards only, changing the OR schedule might not be an option.

9.3.3 Perfomance

When starting a process improvement project, we must first establish the current performance, a so-called zero-measurement. Otherwise, it will not be possible to prove in a later stage that the proposed solutions have indeed led to an improved performance.

The zero-measurement of the current situation is done together with the design of the problem cluster. This requires the definition of KPIs and their targets. Both are established based on stakeholder inquiries. The zero-measurement helps to objectify the perceptions of the problem owner. Note that in order to get a good understanding of the current situation and future scenarios, multiple KPIs should be defined to determine what satisfactory performance entails. When only targeting a single KPI, unrealistic outcomes can be favored over realistic ones. For example, when only targeting the bed occupancy as a performance measure for the wards, it is optimal to reduce the bed capacity to a low level, such that there are always enough patients that need a bed. However, this comes at a cost of a high rejection rate, which is undesirable.

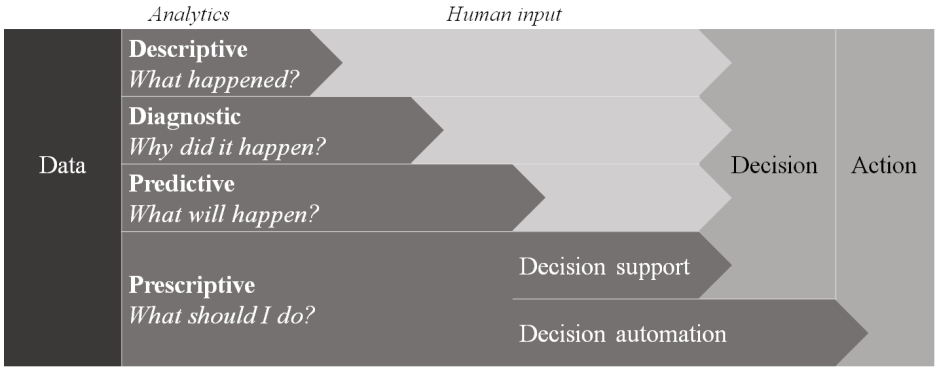
As a patient follows a care chain through the entire healthcare organization, it is important to optimize processes while addressing the entire care process. Therefore, organization-wide KPIs should be developed, besides departmental KPIs. Furthermore, as we analyze processes and the impact of interventions on those processes, not only KPIs regarding the quality of care and care outcomes should be considered, but also KPIs regarding efficiency, effectiveness, and quality of work. This way, processes are designed that excel in all these aspects.

During the prototype phase, when designing solutions, the selected set of KPIs should be taken into account, for example by multi-objective modeling, or by considering the most important KPI as a single objective, while restricting the other KPIs to certain bounds.

As each stakeholder might have his or her own definition of an indicator, each KPI must have an unambiguous definition. There are for example multiple definitions regarding bed occupancy, such as the financial occupancy, which counts how many patients occupied a bed per day, or the operational occupancy, which assesses the percentage of time during the day that a bed was occupied. An ambiguous definition might lead to situations where a bed occupancy of 200% is presented, based on the number of patients that occupied the bed, whereas in the problem context it is important to know that 70% of the day the bed was occupied.

Each KPI should have a target value. This target value shows what is considered a good performance in terms of this KPI. For example, the department can

Figure 9.1 Data analytics capabilities framework (derived from [105])



indicate that they consider the ward performing well when the operational bed occupancy is higher than 80% and a maximum of 5% of the patients is rejected.

In our experience, various departments and organizations experience similar problems. For example the design of an MSS is required in every hospital with a surgical unit. However, for similar problems in various organizations, different solutions might be required, as the ideas about good performance is contained in a different set of KPIs and target values.

9.3.4 Solutions

The design of new planning and control solutions follows from the problem and data analysis. The approach and extent of the data analysis depends on the type of analytics and the amount of human input that is desired in the prototype design, as aptly illustrated by the analytics capabilities framework of Gartner [105] (see Figure 9.1).

The framework contains four stages. First, data has to be transformed into information, before it can be used for decision making. This stage is called *Descriptive* analytics, in which the question 'What happened?' is answered using data analyses, visualization tools, and qualitative methods. When the data-analysis stops here, the project team designs a solution based on the results to mitigate those adverse effects. Although the amount of impact cannot be predicted, and the root cause of the problem might still exist after implementation of the solution, the adverse effects are most likely neutralized.

In the next analytics stage, called *Diagnostic* analytics, the question 'Why did it happen?' is answered using further data analyses and root cause analysis principles (see Section 9.3.2). This way, not only the problem at hand, but also the underlying factors can be tackled. When the data-analysis stops here, the project team designs a solution based on the results to solve the core problem. Although the amount of impact cannot be predicted, the root cause of the problem, and therefore the adverse effects caused by this problem, are neutralized.

The third stage is called *Predictive* analytics, in which the question 'What will happen?' is answered using statistics, data mining, and forecasting techniques. When the data-analysis stops here, the project team can prospectively asses expected outcomes of possible solutions, and can identify future challenges in an early stage. This facilitates better decision making for the project team, as it for example allows selecting to implement the solution with the best expected future performance.

In the final stage, high quality solutions can be obtained by taking on planning and control to influence future outcomes, for example by implementing new planning rules that staff has to follow, or by automating certain decisions. This is called *Prescriptive* analytics, in which the question 'What should I do?' is answered using optimization techniques and simulation to support decision making. (Near-)optimal solutions are designed, and the expected impact in practice of the proposed solution is known.

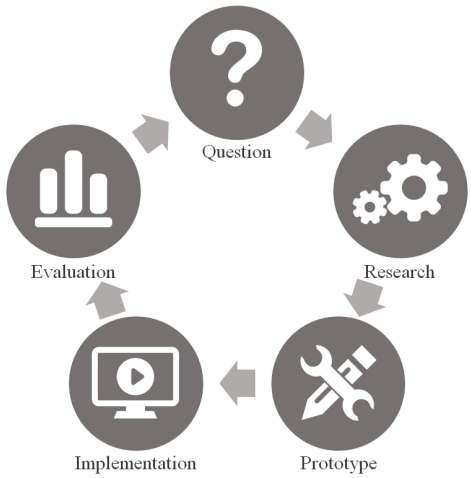
In this final stage of maturity, the expected impact in practice of an automated solution is often the closest to the real impact in practice after implementation, as no human input is required, and full coherence with the decision is reached. When human input is required for decision alignment, the real impact in practice might be less than expected.

As the results from the actions taken give new data inputs to the system, based on which new questions present oneself, this framework behaves as a loop for continuous optimization.

Currently, we see a trend that healthcare organizations start to identify possibilities with respect to higher data analytics levels. Where in the past the available data was mainly used for descriptive and diagnostic purposes, the possibilities that data provides in predictive and prescriptive analytics are being noticed and accepted in healthcare, especially in relation to the operating theater. Examples are the forecasting of beds in the wards based on the surgery schedule, or dividing surgical capacity on a tactical level.

Predictive and prescriptive analytics asks for more involved analytic tools, such as OM/OR techniques, which are especially applicable to prescriptive analyses. Operations researchers can for example design new planning rules and other decision support systems, which can assist or replace current decision makers. Note that when human input is involved, it is crucial that not only the software solution has an easy-to-use user-interface, but also that the employees who are required to work with the solution understand why and how they have to change their current planning behavior. In the descriptive and diagnostic analyses, more human input is involved in the design of a solution. This allows for various types of impact depending on the type of analysis, and various challenges regarding human interplay occur, depending on required analysis for a project. For example, in a diagnostic analysis, operations researchers can act as a sparring partner, and add a systematic and analytical way of reasoning to a project group.

Figure 9.2 CHOIR ecosystem



9.4 The ecosystem of education, research and impact

Now the theoretical background of process design, improvement, and optimization is known, this section presents the network of healthcare institutions with whom we facilitate the CHOIR ecosystem of education, research, and impact. First, we explain the ecosystem in Section 9.4.1, together with the network of hospitals and other involved stakeholders in Section 9.4.2. We furthermore discuss the three main activities of CHOIR: education, research, and impact, in Sections 9.4.3, 9.4.4, and 9.4.5 respectively.

9.4.1 The ecosystem

Projects executed within the CHOIR ecosystem consist of five phases, as shown in Figure 9.2. Currently, CHOIR's predominant involvement is with these first two phases of the ecosystem. However, after a successful research project, impact in practice is not necessarily derived. Therefore, the ecosystem should be continued with a prototype, implementation, and evaluation phase.

Each project starts with a *question*, originating in healthcare practice. We stipulate involvement of clinical staff in every project, to prevent not having their support from the offset. In many cases, an organization or department approaches us with a question, which is often formulated as a solution. This solution is regularly, as mentioned before, in the form of more capacity. For example: 'The work pressure for our staff is way too high, can you calculate how much more capacity we need?' Most questions are based on perceived issues, such as a high workload in our example. However, the human mind tends to remind those incidental situations over common situations in which no pressure was

present. Furthermore, an objective interpretation of these questions is regularly not available, as performance is not quantitatively measured, but subjectively interpreted.

To address a question from practice in a systematic way and to find the root cause problem, we first determine the core problem(s), as explained in Section 9.3.2. Furthermore, we analyze the initial performance of the system using quantitative and qualitative analyses techniques, as explained in Section 9.3.3. This is required to objectify the perceived problems, to establish whether any operational data is present, and in order to establish after the project whether performance has improved. For this purpose, we define KPIs and gather target performance levels from administrators.

Using these systematic analyses, the perceived problem becomes a quantified and objectified problem. This regularly causes a change in research question as well. Continuing the example, the analysis might have shown that the work pressure for staff is indeed 20% above target in the mornings, but below target in the afternoon. The research question then becomes: 'How to allocate work over the day, to level the work pressure for staff?', or 'How to better align the staff schedules to the workload over the day?'

After the question is well-designed, the research goals are clear, and the target performance is known, *research* is done. This starts with a (literature) search for approaches that can overcome the gap between the current and the desired situation. After this search, solutions are generated using a systematic solution design process, or tooling can be developed to design a(n optimal) solution, as explained in more detail in Section 9.3. Tools that visualize (redesigned) processes can provide much insight to practitioners, and lower the barrier for acceptance. Particularly computer simulation can greatly assist to demonstrate the expected performance of the solution in practice and to convince healthcare employees of the solution's impact in practice.

After a valid solution has been chosen, a *prototype* is developed and/or a pilot is started. A prototype is developed when software is required to assist in decision-making. A pilot is started to evaluate the use of the new planning rules or decision support software in practice. A pilot is always conducted in a small setting, in order to evaluate the requirements of the implementation phase, to analyze the needed support during a full implementation, and to analyze and overcome the shortcomings of the solution in practice.

A successful prototype and pilot is followed by a full *implementation* of the solution. This not only includes a well-designed implementation plan and support during the implementation, but also aftercare and continuous development of the tooling to support the needs of practice.

After the implementation of the solution in practice, an *evaluation* takes place to empirically assess whether the solution resulted in the expected performance improvement. From the evaluation, new questions come up, which makes this five phases act as a cycle.

9.4.2 Stakeholders

Depending on the phase of the cycle, various stakeholders are involved to enable a successful impact in practice. Among these are the patients, front-line staff, healthcare administrators, business partners, bachelor, master and PhD-students, and faculty members.

Gathering patient preferences is essential input for system redesign, but is rarely reported in OM publications. Patient inquiries, for example, using discrete choice experiments, may uncover operational constraints and performance indicators.

Care professionals are involved in all phases of the cycle, as the project is executed within the healthcare institute. Our network of care professionals is involved after a first implementation and evaluation cycle in the hospital under study, to show the results in practice, and to enable further dissemination of the results among other healthcare institutes.

Business partners play an important role in the implementation phase, by providing a complete business solution based on the research and prototype. This includes the development of a reliable software tool after prototyping, and customer support during and after the implementation. They also take part in the evaluation phase, to evaluate effectiveness of the approach and improve upon reported issues. A fruitful relation with such a partner is potentially mutually beneficial, as well-implemented research leads to new research questions, and more opportunities for implementation. Also, it may generate revenues, which preferably would (partly) flow back to fund research. For this reason, in 2014 CHOIR has started a spin-off company, called Rhythm BV. By collaborating with this spin-off, we may use part of the income of the partner for new research opportunities. The same can be obtained through a royalty or intellectual property contract with an external partner.

Students are a main driver of our research projects. We find it essential for CHOIR researchers to be present in healthcare organizations, to lower the barrier for practitioners to approach us, to ensure relevant topics of study, and to promote involvement of front-line staff. Therefore, we strongly believe that researchers should be positioned in a hospital for a major part of their research, ideally to be considered by practitioners as part of their own organization. Therefore, we follow the researcher-in-residence model in all stages of the research project. The researcher-in-residence model positions a researcher as a core member within a healthcare team of relevant care professionals [194]. In a context of process improvement, the researcher brings a new body of expertise, focused on data-analysis, modeling, and structured decision-making, which is different from, but complementary to, the expertise of the existing team. Within CHOIR, we differentiate between three types of student-researchers, with their own skill set. Bachelor students excel in descriptive and diagnostic activities. Within a project team, they perform a thorough problem definition, together with a root cause analysis. Based on this analysis, they present improvement opportunities to assist decision makers. Master students not only assist in descriptive and

diagnostic activities, but can also take on a predictive or prescriptive aspect to enhance decision-making support, depending on the organization's need. PhD students collaborate in projects spanning all possible activities, including descriptive, predictive and prescriptive activities. It is beneficial to include PhD students in projects from an early phase, as a good understanding of the problem and its causes, and the perception by clinicians and other healthcare staff that an outside person is fully aware of all restrictions that apply, leads to high acceptance and cooperation rates, and therefore a better performance of the proposed solutions. Furthermore, as PhD students reside in a healthcare organization for a longer period of time, pilot and early implementation results can be evaluated to prove impact in practice not only analytically, but also in an empirical way.

Finally, faculty members are important contributors to the CHOIR ecosystem. Although they regularly visit healthcare institutes, they do not participate in the researcher-in-residence model themselves. This prevents intellectual isolation of their junior staff and students, which is a common disadvantage of the researcher-in-residence model [194]. Being a researcher-in-residence requires great professional skills, which are best trained on site. Also, evidently it requires academic skills, which are best trained at the university, with faculty members and fellow researchers. The CHOIR researchers and PhD students therefore reside at the university together for at least two days per week for this purpose, and to share experiences and collaborate with other researchers-in-residence. Faculty members not only educate students and care professionals, but also monitor the applied methodologies in the various projects, connect healthcare institutes with similar research questions, and link relevant people within the CHOIR network. Through their diverse activities, faculty members are involved in all three pillars of CHOIR (education, research, impact), which we elaborate upon in the upcoming sections.

9.4.3 Education

The first pillar of CHOIR is education. By educating both students and healthcare professionals, CHOIR aims to bridge the gap between theory and practice.

As a research center within the University of Twente, CHOIR's main educational focus is on BSc and MSc students of Industrial Engineering and Management, Health Sciences, and Applied Mathematics. The teaching activities encompass OM/OR in healthcare through lectures and practical sessions, but also professional and academic skills training and practical experience, through graduation projects and internships. We aim to educate students to become independent researchers within an organization that speaks a different language. After their studies, many CHOIR alumni continue to spread the knowledge throughout the organizations they work in, which results in a growing healthcare logistics community. Indirectly, these alumni make the greatest impact of CHOIR's activities.

Aside from academic students, CHOIR also educates healthcare professionals, including managers, administrators, logistics staff, doctors, and (head) nurses.

Most of them have an educational background in medicine or nursing, and lack OM training. From experience, these professionals often know the practical constraints of process optimization, but lack methodological knowledge and knowledge about theoretical (im)possibilities. Also, they typically find it hard to look at operational processes in an integrated way. Instead, they tend to only focus at their department or role in the system. The course encompasses not only theory. Parallel to the course, the participants have to perform a process improvement assignment within their own department, under supervision of CHOIR staff. In our experience, many former participants in the course become champions for our research. They often initiate new research projects, and serve as in-company supervisors for our student projects.

9.4.4 Research

The second pillar of CHOIR is research. We take on complex logistical challenges that are driven by practice to design or optimize the organization of healthcare processes. Herein, we aim to improve the quality of care, the quality of labor, and the efficiency of processes. We find it important to emphasize this, as process optimization is quickly solely associated with efficiency and working harder.

As it is evidently undesirable to try out interventions in practice, in our research we make use of mathematical models and computer simulation to prospectively asses the performance of an intervention before actual implementation.

We disseminate our research in the scientific community through two main channels. First, we present and publish our results in the OM/OR domain, where field experts can give us feedback on the methods used. Second, we present and publish our results in the medical domain, to show the potential of the use of OM/OR tools for optimizing healthcare processes.

CHOIR receives its funding for research and PhD projects both from practice, e.g., through funding from hospitals, and from funding agencies, e.g., through national science programs.

9.4.5 Impact

The focus of our education and research is to have an impact in practice, which is the third pillar of CHOIR. For a knowledge gathering and developing center such as CHOIR, impact is the dissemination and effective application of that knowledge in practice.

We disseminate knowledge through the network of healthcare providers by the positioning of our students in healthcare organizations, through seminars and symposia at our university and at hospitals, through teaching, through alumni, and through publications in professional and academic journals.

We apply knowledge in practice in various ways. Through the projects of bachelor, master, and PhD students we contribute to process optimization in practice. Although not every project results in implementation, the presence of someone with a different background that questions regular protocols raises an

awareness for improvement potential in organizations. Also our spin-off makes an impact in practice through various activities, such as (analytics/decision support) tool development, training, and change management. Finally, the greatest impact is made by alumni who remain active in the healthcare sector.

In the next section, conditions for impact will be discussed, based on our experiences with the CHOIR ecosystem related to the work in this thesis.

9.5 Conditions for impact

In this section we reflect on the impact we have made in practice with the projects related to this thesis, and analyze the critical success factors. We discuss seven deliberations that we encounter in our research projects, which we illustrate using examples of student projects that were executed in UMC Utrecht as well as the projects of this thesis. The deliberations are:

1. risk-adverse approach vs. engineering approach (Section 9.5.1),
2. bottom-up vs. top-down (Section 9.5.2),
3. theoretical projects vs. practical projects (Section 9.5.3),
4. theoretical solutions vs. solutions from practice (Section 9.5.4),
5. productivity vs. service level performance (Section 9.5.5),
6. individual champions vs. champion organizations (Section 9.5.6), and
7. decision support vs. decision making (Section 9.5.7).

9.5.1 Risk-adverse approach vs. engineering approach

Healthcare practitioners are inclined to be risk-adverse with every challenge they encounter. This is required in their evidence-based clinical practice, where they prescribe treatment or medication for a patient based on proven risks and effectiveness. Consequently, when optimizing processes, practitioners are inclined to copy better practices. On the contrary, engineers are educated to solve problems by designing entirely new solutions, and by performing experiments. These experiments are not performed in practice. Instead, engineers use mathematical and computer simulation models to prospectively assess the impact of various alternative solutions. For engineers, experimentation is about learning and understanding the system, while for practitioners experimentation is limited and bounded, as it may be harmful for the patient.

When operations researchers start working with healthcare practitioners, it is important for both parties to be aware of this different way of thinking. In our experience, practitioners are often appreciative of (simulation) models. Particularly if a model visualizes the situation after the intervention, this creates evidence for practitioners of the effectiveness of the solution.

The ophthalmology emergency surgery scheduling project is an example of a successful collaboration of a project team with people from various backgrounds, including ophthalmologists. The engineering approach was followed by [138], who joined the project team, and explained all steps he took in his research, from data analysis, to simulation building, to experimenting. The simulation study aided in the acceptance of the results by the evidence-based orientated care professionals. This way, the care professionals were aware of the approach taken, and the recommendation was easily accepted by the involved ophthalmology staff.

Sometimes, our students only collaborate with staff that have not had any training in medicine. The facility layout project of [22] is an example of a project where all involved hospital stakeholders already understood the engineering approach, as only staff were involved in the decision making process who had previous experience with facility layout problems. Preceding the redesign of the wards in UMC Utrecht, [22] analyzed the walking behavior of nurses, in order to propose a new layout of the wards that minimizes their walking distance.

9.5.2 Bottom-up vs. top-down

In Section 9.2 the bottom-up and top-down approaches were discussed. As argued, the joint approach gives the highest probability of success.

When top-down support is present, but problems are not perceived by front-line staff, implementation possibilities of project results are minimal. Van Sark [270] analyzed the multi-disciplinary way of working at UMC Utrecht's endocrine oncology department. Although she was able to develop a simulation model that captured the department's employees behavior, and showed that the implementation of new planning rules could result in improved performance of the department, no follow-up implementation initiatives were taken based on this project. One of the reasons was that the employees already met their performance targets, and did not feel the need to change their current behavior to possibly further improve upon their performance.

Another example shows that when bottom-up support is present, but top-down support is lacking, implementation of results is challenging as well. In the WKZ (in Dutch: Wilhelmina Kinder Ziekenhuis, UMC Utrecht's child hospital) an exhaustive renovation is scheduled for the near future. Therefore, design plans are developed for the operating rooms, intensive care unit (ICU), daycare and wards. One of the questions withing the building program was how many neonatal and pediatric ICU beds were required, when in the new setting those departments join forces instead of working separately [226]. However, the real interest of management was in how to assign a fixed number of 67 beds to the two units, and how many beds should be flexibly allocated on a day-to-day level to the department with the highest needs. Although great bottom-up support was present during this project, top-down support was initially lacking, making it challenging to find the right research question and the boundaries of the project. Later on, this support was present, which influences the likelihood of having long-term impact.

The research project of Chapter 3 and Chapter 4 at UMC Utrecht's pathology department was initially top-down initiated. To deliver a successful project, much effort was needed into getting the front-line employees involved. Being a researcher-in-residence was helpful. By incorporating the front-line employees into the problem finding and solution design phases, they supported the final recommendations of the project, and were eager to implement the new planning rules in practice.

Recently, capacity managers in UMC Utrecht started to focus on care chain optimization. Through several research projects, the relationship of the OR, ICU, wards and day care units are analyzed (e.g., [284]). However, care chain optimization implies that the care chain as a whole is optimized, which might result in improved overall performance at the expense of reduced performance for single chains. Results showed, for example, that a new MSS can reduce the variation in bed occupation of multiple wards. However, when top-down support is lacking, for example by only evaluating KPIs on a departmental level, or by not including the global management level in the project team, the analyses will not result in impact in practice, despite the improvement possibilities.

9.5.3 Theoretical projects vs. practical projects

The theoretical requirements for scientific research projects can be demanding. For example, modeling is one of the requirements for our BSc and MSc students Industrial Engineering and Management and Applied Mathematics to successfully finish their thesis projects. However, this theoretical requirement may contradict with the needs in practice, where a straightforward and easy-to-implement solution can have great impact already. Furthermore, there are projects with high potential for theoretical contributions to the scientific community, which are not relevant to practice. On the contrary, there are also projects with high practical relevance, but with little significant contribution to science.

For all our research projects we always aim to combine a theoretical and practical perspective. For example an analysis of the surgery schedule of the oncology department of UMC Utrecht, was initially started from a practical perspective. An interesting scientific question regarding the concept of alternating emergency operating rooms, was combined with the original questions from practice in a simulation study. Unfortunately, this project showed no immediate gains were available in the UMC Utrecht case, due to the tight surgery schedule during the rebuilding phase of the operating theater [281].

Concluding, the difference between theoretical projects and practical projects might lead to skewed expectations from people involved. Research projects are more likely to have a longer throughput time compared to improvement projects, as certain requirements have to be fulfilled. Furthermore, they do not necessarily lead to improvements, as favorable results are not guaranteed in research. When research projects are executed, they can have a theoretical focus, or a more practical perspective. However, both aspects should be included in some way to ensure a successful project for both the researcher and the organization.

9.5.4 Theoretical solutions vs. solutions from practice

Inundated with operational problems, managers are inclined to solve the problems at hand. When problems increase in frequency or size, we find that problem owners rapidly advocate the necessity of more capacity as the solution. However, in our experience with improvement projects in healthcare settings, in hardly a handful of cases there was a proven capacity shortage. Nevertheless, increasingly, healthcare providers realize that the rising expenditures need to be countered, and that more capacity is no longer an option. Instead, new process designs, and new planning and control models are sought after to overcome their challenges. As mentioned in Section 9.4.1, the improvement question is then often formulated as a solution.

After identifying the core problem, as explained in Section 9.3.2, we regularly find ourselves evaluating the proposed solution as a possibility to solve the problem. This is a practical perspective, wherein the possibilities for implementation are considered more important than finding the optimal solution. Note that in practice, a straightforward, near-optimal, planning solution most likely leads to higher impact than an optimal solution, as employees adhere better to easier-to-understand planning solutions than to more complex solutions.

From a theoretical OM/OR perspective, exact methods to find or design an optimal solution are preferred over evaluation studies. From a practical perspective, evaluation studies are preferred, for example using computer simulation. Exact methods enable the researcher to determine the best possible decision for the project team. As the decision is optimal, it shows the best possible performance that can be reached, which can serve as a benchmark to the organization's performance. On the other hand, evaluation studies enable a researcher to test several interventions and scenarios. They give the involved healthcare staff more flexibility in testing those interventions which they consider promising to implement in practice. As evaluation studies often include a visual component, they are also relatively easy to follow and understand, which supports the acceptance of final recommendations even more.

In our experience, most successful studies from which one or more recommendations were implemented, involved some kind of evaluation component. For example the histopathology process improvement project of Chapter 3 and Chapter 4, where several possible interventions were evaluated. As decision makers were involved in the design of the interventions themselves, they know whether implementation of the intervention was realistic and possible. From a research perspective, it might be interesting to add several theoretically interesting solutions, or alternative solutions, for example to serve as benchmarks.

Complex solutions, which are for example the outcome of an optimization study, have a higher probability of implementation when their implementation depends less on human input. An example is the schedule template in the outpatient clinics' agendas of Chapter 6. Based on this research, a solution was programmed into the computer system used for appointment planning, which reduces the possibilities of misuse and non-adherence.

9.5.5 Productivity vs. service level performance

Most of CHOIR's projects revolve around problems related to service level performance problems, such as excessive waiting or access times for patients, low resource utilization, or highly fluctuating workload levels. While hospital administrators tend to be inclined to maximize utilization of their costly resources, they often do not realize that high utilization combined with variability results in strongly fluctuating workloads, as well as waiting and access times. For example, as the operating room department has typically been designated as leading in care pathways, a high operating room department utilization is strived for. In combination with variability in operating room supply and demand, this results in bullwhip effects in up- and downstream departments, such as the adverse effect of high patient waiting and access times, and highly fluctuating workloads for staff [296]. While these are elementary queueing theory principles for OM/OR scholars, we have experienced that this is perhaps the biggest eye-opener for practitioners.

Since variability affects operational performance, in our solution approach we always first strive to reduce variability. While there is natural variability, much of the variability in care processes is artificial, i.e., caused by the organization itself. For example rigid block schedules and capacity allocations in outpatient clinics, radiology, and operating rooms, result in strongly fluctuating workloads. Another example is a low adherence to care protocols, or to plans or schedules. As a next step, we try to forecast the remaining variability. As argued before, better forecasts create the potential to increase operational performance. For the remaining variability, flexibility is required to alleviate its effect. For example, flexible capacity can be used to deal with sudden fluctuations in demand. When redesigning the planning and control of processes, we look for opportunities to create such flexibility, which we can use as buffer to alleviate the effects of variability and thus improve operational performance. On higher planning levels, there is potentially greater flexibility, albeit with more demand uncertainties due to the longer planning horizons. Since capacity availability (a strategic level problem) is most often sufficient, and there is ample effort to improve operational level performance, much of our research addresses the tactical planning level. This planning level is often overlooked or deemed too complex by practitioners [145]. However, if underexposed, not utilizing the inherent flexibilities on the tactical planning level results in reduced operational performance.

Flexibility, unfortunately, often comes at greater operational costs, thus affecting productivity. For example, more flexible staff require more training, open access systems may result in lower utilization than appointment systems, and overtime and surplus capacity are costly. As a result, while designing planning and control, we need to make a trade-off between service level performance (i.e., maximize the quality of service and quality of labor by managing variability) and productivity (i.e., minimize the costs incurred by creating flexibility).

In UMC Utrecht we have experienced this trade-off in multiple (student) projects. For example when analyzing wards or outpatient clinics, where utiliza-

tion is a frequently studied KPI [49, 226]. However, a target utilization level has consequences for the number of rejections at the ward and cancellations and waiting time at the outpatient clinic, for which a target performance was not discussed before. In these projects, we were able to show that when the variability in patient arrivals, patient length of stay, or service duration was reduced, not only the service levels could be improved, but also the productivity. Possible interventions included re-assigning the specialties to wards, pooling of resources, and new agenda templates.

9.5.6 Individual champions vs. champion organizations

Research projects are not only successful due to the researcher, but to a large extent also due to the project team surrounding the researcher. Many of the hospitals that CHOIR collaborates with, have a so-called champion. This champion supports the CHOIR ecosystem, and is the connector between CHOIR researchers and the hospital administrators. Besides champion individuals, also champion departments or organizations are present. To benefit from process improvement projects, it is essential to grow from a champion individual to a champion department or organization to ensure the continuity of process improvement in hospitals. This way, the implementation and continuation of the results of a project does not depend on a single individual.

Collaborating with champion institutions allows for long-term relationships, which enables to take on larger research projects with long term commitments. These projects typically can result in more impact, as not only quick wins are derived, and the project scope can be extended over multiple departments and hierarchical levels.

Not only for the collaboration of CHOIR with healthcare organizations, but also for process improvement projects within a department there is a risk of collaborating with champion individuals, as the continuity of the improvement project depends on one individual. However, those individuals are important in initiating projects. For example, the research project in the department of Pathology was initiated by a champion individual (see Chapter 3 and Chapter 4). As the recommendations from the project do not only involve changed behavior from a single individual, but from all staff members, it was important to involve these other stakeholders in the project as well, to ensure they would give the required input for the project, and comply with the recommended change in behavior. The recommendations consisted of simple planning rules, which were easily adopted by the employees as they matched their perceptions. When champion individuals are developed into champion departments or organizations, process improvement projects can really take off. An example is the multi-disciplinary clinic project of Chapter 6. We were added to the project team for quantitative support from the start of the project. This way, we were able to support in many ways, for example with the capacity requirements and the design of template schedules.

9.5.7 Decision support vs. decision making

Key in successful decision making for an organization is a clear vision. Based on this vision the decision makers can evaluate several possible interventions for improvement, with corresponding expected gains and other consequences. However, in practice, this evaluation is more involved. The overall vision needs to be translated in clear performance measures and target KPIs, as shown in Section 9.3.3. As every stakeholder gives different weights to the importance of the various selected KPIs, there may be several optimal solutions, depending on the targeted stakeholder.

The task of CHOIR researchers is to provide decision support, not to be the decision maker. This asks for an integrated approach, in which close collaboration with the stakeholders, for example the project team members, is sought. OM/OR tools can evaluate solutions, and assist in showing the consequences of certain decisions in terms of the selected KPIs, for example using predictive analytics tools. Furthermore, OM/OR tools can provide optimal solutions, which an organization can choose to implement, based on prescriptive analytics tools.

When supporting a multi-disciplinary project team, it is important to determine the KPIs together with all disciplines. During the analysis of the required ICU capacity for WKZ when using shared capacity between the two wards, the neonatal ICU and pediatric ICU both had their own definitions of certain KPIs, such as bed occupancy [226]. Furthermore, the available data for both departments differed, which makes a fair comparison between the two departments hard. However, through communicating these differences, and by involving the decision makers in the analytics process, the results of this study were accepted and used in the WKZ building process decision making.

9.6 Conclusions and discussion

CHOIR aims to make an impact in practice using a scientific approach. However, not every project comes to its full potential. Based on our experience, we identified multiple conditions for impact. They are not restrictive, but indicate whether projects are more likely to have an impact in practice.

First, a project can only be successful when scientific and healthcare people form a project team together. This means that scientific staff is introduced to the healthcare environment (researcher-in-residence), and that healthcare staff is introduced to the engineering approach. When cooperating with clinical staff as an engineer, the engineering approach should be clearly explained. This explanation should be based on concepts that are familiar to them, to get them involved in the project, and to get them to trust the methods used. This is important, as after the prototype phase they have to take the lead in implementation of the solution. When they do not fully trust the predicted impact this solution will have in practice, the chances of actual implementation are minimal.

Second, in line with the first condition, ensure that projects do not rely on a single person in the healthcare organization, but get the whole department

or people from within the whole organization involved. This ensures that after the prototype phase, multiple people continue with the ideas resulting from the project.

Third, combine top-down and bottom-up approaches. This way, there is pressure to change the current situation, and there is commitment for the project from the front-line staff. From a bottom-up perspective, it is especially important that the management level is involved in the project, as their support ensures that time and money is freed when necessary, and decisions are taken. Note that in care chain optimization, this means that not only the managerial level of the individual chains should be involved, but also a global level manager.

Fourth, a project should be a balance between theory and practice. The systematic problem solving approach adds value by not solving consequences, but the root cause of the problems. Furthermore, OM/OR approaches help organizations to improve their organization of processes using theoretically sound solutions. On the other hand, keep in mind that when decisions in hospitals involve many people, solutions should be easy to implement. In automated systems, more involved solutions are appropriate.

Fifth, clearly distinguish that researchers support in decision making, and end users/management are the decision makers. This ensures that end users are involved, and increases the probability that real impact is made.

Sixth, define and measure at the start of a project measurable ambitions in terms of quality of care, quality of work, and productivity. Through these measures, the effects after implementation of research results in practice can be measured, which shows whether the intervention has been successful.

Despite these conditions, creating a sustainable impact in practice with process improvement (research) projects is still challenging. We discuss two challenges from a change management point of view and OM/OR point of view.

From a change management point of view, there are multiple aspects that come in mind with regard to implementation of process improvement solutions. Despite having fulfilled all aforementioned conditions, implementations can still fail. Aspects such as the institutions culture towards change, the degree of urgency, the knowledge and skill level of involved staff, and the organizational structure (e.g., centralized or decentralized) have a major impact on the researchers possibilities to make an impact in practice. Although researchers cannot interfere with most of these aspects, it is important to keep them in mind when starting a process improvement trajectory.

From an OM/OR point of view, most of the scientific challenging research projects are at the tactical level of healthcare planning and control, and most impact can be derived with tactical planning solutions. However, the tactical level is at the same time the level that healthcare practitioners are most struggling with. First of all, to benefit from tactical level interventions, larger interventions, such as the introduction of flexible capacity, are often required. Despite the benefits, it is hard for hospitals to motivate employees to share their capacity in order to flexibly allocate it to those who are in need of some extra capacity. Furthermore, it is challenging to determine the right interval at which the tactical

decisions need to be made, as it is another illustration of the trade-off between the need for variability reduction and complexity reduction. On the one hand you want to be as flexible as possible, which suggest a late assignment of capacity. On the other hand, employees require timely clarity about their rosters, which is formalized in collective agreements.

CHAPTER 10

Outlook

High quality cancer care not only requires outstanding medical expertise, but also an outstanding healthcare infrastructure. Patients follow immensely diverse care trajectories, causing healthcare processes to be very interrelated. The design of healthcare processes is therefore of great importance. A well organized hospital enables patients to receive the care they need without unnecessary waiting. It enables practitioners to deliver care to as many patients as possible within the limited time they have, without the need to focus on logistical details, but focus on the patients' needs instead. It also enables hospitals to stay financially healthy, by efficiently using their resources.

Healthcare process optimization and design requires an integrated, system-wide approach, as Part I of this thesis explains. A review of the current literature on integrated multi-disciplinary care processes in Chapter 2 shows the current state-of the art, and identifies the open research questions in this field. We outline that there is a need for robust planning solutions that can deal with the inherent variability of complex multi-disciplinary systems. After a general introduction, Part II, Part III and Part IV address various stages in a cancer patient's journey. Trade-offs in the design of each of these stages are made explicit, to support decision makers with insights into the consequences of possible interventions. Consider as an example appointment planning, where patient behavior constrains the possible productivity of clinics (Chapter 5), and where a shared resource needs to be used efficiently, but also equitable (Chapter 6). Part V discusses the realization of mathematical results in healthcare practice (Chapter 9). As we aim to support decision makers, a multi-disciplinary collaboration is required for impact in practice. This concluding chapter provides an outlook on integrated process optimization in cancer care. We discuss four trends in cancer care:

1. Multi-disciplinary care
2. Shared resources
3. Personalized care
4. Centralization of cancer care

Trend 1: Multi-disciplinary care Through specialization, medical specialists are increasingly becoming experts in a specific disease area, whereas patients

are increasingly co-morbid, as various cancer types have become a chronic disease. These two trends require multi-disciplinary care approaches, in which multiple specialists from various disciplines work together and share their knowledge in order to determine what treatment strategy works best for their patient. Also, as seen in Chapter 1, para-medical staff are more often included in the treatment design of patients, which requires additional planning as well. Furthermore, as cancer has become a chronic disease, to a large extent care will be delivered by care professionals outside a hospital. Therefore, the integration of hospital care and 2nd line professionals, such as general practitioners (GPs) and nursing homes, is of great relevance, and which is currently underexposed. Jointly optimizing the processes from 1st and 2nd line care professionals from multiple organizations will provide efficient and effective integrated cancer care processes.

Studies on the organization and implementation of multi-disciplinary care planning are scarce, as only recently hospitals are faced with larger patient populations that require well-organized and coordinated multi-disciplinary care. Furthermore, the organization of multi-disciplinary care clinics comes with its own challenges, such as aligning staffing schedules over departments. This is something that we did not consider in the work of this thesis, and is subject for further research.

Trend 2: Shared resources As multi-disciplinary care is becoming the standard, more specialists are involved in the treatment of (cancer) patients. This not only involves medical staff, but also para-medical staff are partaking in the treatment of patients, as we showed in Chapter 3, 4 and 6. These specialists do not solely focus on cancer patients, but also serve patients with other disease types. In order to provide a high quality of care to all patients, not only cancer care processes should be taken into account when optimizing cancer care chains, but also the effects on the remaining patient population that uses those shared resources. There are multiple ways to integrate shared resources. First, the regular use of the resource can be preempted in order to fulfill the need for the specific care. An example of such usage of a shared resource is the pathology laboratory, where cancer tissue was prioritized over non-cancer tissues, in order to deliver a rapid diagnosis to the patient (see Chapter 3 and Chapter 4). Second, a resource can be partially dedicated to the specific care. An example of such usage are rapid diagnostics trajectories, where on specific days of the week X-ray slots are reserved in the radiology department for possible cancer patients. Third, the regular use and the use for specific care of the resource can be integrated, where they both use the same resource together. An example is the multi-disciplinary clinic, where both patients with regular care demand as well as cancer patients are seen (see Chapter 6). In practice, it is hard to truly integrate care on a shared resource, as practitioners are afraid that other users take advantage of their capacity. Therefore, the impact of this thesis on shared resource usage beyond cancer disciplines is minimal. However, research has shown that when dedicating capacity to specific patient populations, the remaining care is disadvantaged [297]. Therefore, it is important for OM/OR researchers to show

the side-effects of interventions incorporating shared resources, and to show that integrated care is beneficial for the overall patient population.

Trend 3: Personalized care Technological advances, such as DNA sequencing methods, enable patients to more specifically analyze the type of cancer they suffer from, as well as the type of medication that will be most effective. When these methods become more commonly used in cancer diagnostics and treatment, cancer care will shift from standard care pathways towards personalized care pathways. In line with the multi-disciplinary care trend this causes many deviations from the 'standard' pathways, which results in many exceptions. This requires a new planning approach, which is robust for a variety of diagnostics and treatment options depending on the patient's need. Furthermore, as exceptions are in general seen as disturbing the processes, and should be reduced according to the principle of Operations Management (OM) that minimizes variability and complexity, one should strive for a similar vision as the pathology laboratory (Chapter 4), who state: 'Rapid diagnostics should not be the exception, but the standard'. Despite this upcoming trend, currently, most applications of Operations Management/Operations Research (OM/OR) in healthcare use the concept on clinical pathways in their models, without for example incorporating variability in appointment sequences (see Chapter 2).

Trend 4: Centralization of cancer care In the Netherlands, cancer care becomes more and more centralized. This is the consequence of the high specialization of care professionals, the technological advancements, in combination with the wide range of possible types of cancer. To ensure that a patient gets high quality care from experienced staff, volume targets are set for healthcare institutes, which forces hospitals to join forces in treating specific types of cancer patients.

A recent challenge for UMC Utrecht is the regional case mix planning. Hereby, multiple regional hospitals collaborate on specific types of cancer care, and patients with specific conditions and diseases are referred to one of the collaborating hospitals, in exchange for patients from another category to be referred back. Case mix planning is a frequently studied topic in for example general practitioners' settings, for example by deciding whether a patient should be admitted into the patient panel of a GP, considering the patient characteristics and future demand. Similar analyses are required for the distribution of cancer patients over healthcare institutes, considering the trends in patient volumes for various tumor types. For example, the incidence of esophageal tumors is growing, whereas the amount of lung tumors is decreasing. Exchanging these patients might seem a good fit for the coming year, but it might be an unattractive decision for the future. Furthermore, analyzing the effects of these patient exchanges requires an integrated view. On a managerial level, financial rewards are of consideration, as treating esophageal cancer patients might be more or less beneficial than lung patients based on the agreements with insurers. From a logistical point of view, the exchanges have an effect on the outpatient clinic capacity, as more esophageal

clinicians are required instead of lung specialists. Further in the care chain these exchanges affect the operating room schedule and the ward planning as well, as every patient category comes with their own expected surgery duration and expected length of stay. Therefore, the whole care chain should be taken into account when making these decisions.

Impact The four mentioned trends in cancer care ask for integrated processes and advanced logistics. This thesis showed that OM/OR can aid decision makers in the design, planning and control of these processes. However, as shown in Chapter 2 and Chapter 9, there is only limited reported impact of the scientific research in this area. To increase the impact of scientific OM/OR research in practice, we observe a need for the following three items in the OM/OR community:

First, the theoretical focus of the scientific OM/OR community needs to be amplified with a focus on practice. Therefore, in the literature, more case studies should be reported, along which topics such as collaborations with practitioners, how these collaborations were initiated, difficulties that were faced when working with practice, and included conditions in the research design to ensure impact in practice.

Second, as the availability of data is a challenge for many OM/OR researchers, we should strengthen the OM/OR community by sharing de-identified real-life data, such as the surgery scheduling benchmark set of Chapter 8. Independent online platforms may serve as repositories for such data sets. These data sets not only stimulates researchers to take on new projects, but also enables a fair comparison of theoretical approaches.

Third, where in the medical literature many real-life results of pilot studies are presented, the OM/OR literature rarely reports on the results of implementation. We should support researchers to not only present their predicted results, but also the results in real-life after implementation. This will give a better understanding of how to make an impact, and an assessment of the actual impact that is made with OM/OR approaches, the challenges that occur when implementing a stylized solution in practice, the consequences for the (near)optimal solution, and relevant conditions that should be taken into account in future research projects to ensure a good outcome in practice.

Conclusions The design and optimization of health care processes in general, and cancer care processes in special, requires an integrated approach, in which OM/OR researchers join forces with health care practitioners to make an impact in practice. This enables more patients to receive the care they need, whilst the quality of care and the quality of work are increased. Furthermore, this ensures that hospitals' processes are robust against the upcoming future trends in oncology, which include specialization, personalization, and centralization. Incorporating multiple departments into the optimization of processes, reduces the risk of local process optimization, diverts possible negative side-effects for other patient groups, and increases the impact in practice. The research presented in

this thesis share this aim, and makes a first step in this direction, as the studies in this thesis show how quality of care and work together with productivity can be jointly optimized in complex, often multi-disciplinary, care settings.

Bibliography

[1] MK Agrawal, SE Elmaghraby, and WS Herroelen. Dagen: A generator of testsets for project activity nets. *European Journal of Operational Research*, 90(2):376–382, 1996.

[2] A Ahmadi-Javid, Z Jalali, and KJ Klassen. Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research*, 2016.

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

[4] S Ahmed and A Shapiro. The sample average approximation method for stochastic programs with integer recourse. *Technical Report, School of Industrial and Systems Engineering, Georgia Institute of Technology*, 2002.

[5] E Alfonso, X Xie, V Augusto, and O Garraud. Modeling and simulation of blood collection systems. *Health Care Management Science*, 15(1):63–78, 2012.

[6] H Allaoui and A Artiba. Integrating simulation and optimization to schedule a hybrid flow shop with maintenance constraints. *Computers and Industrial Engineering*, 47(4):431–450, 2004. ISSN 0360-8352.

[7] H Allaoui and A Artiba. *Hybrid flow shop scheduling with availability constraints*, pages 277–299. Springer, 2014. ISBN 1461490553.

[8] VL Allgar and RD Neal. Delays in the diagnosis of six cancers: analysis of data from the National Survey of NHS Patients: Cancer. *British Journal of Cancer*, 92(11):1959–1970, 2005.

[9] DM Almog, JA Devries, JA Borrelli, and DT Kopycka-Kedzierawski. The reduction of broken appointment rates through an automated appointment confirmation system. *Journal of Dental Education*, 67(9):1016–1022, 2003.

[10] JP Ambuel, J Cebulla, N Watt, and DP Crowne. Urgency as a factor in clinic attendance. *American Journal of Diseases of Children*, 108(4):394–398, 1964.

[11] MR Amin-Naseri and MA Beheshti-Nia. Hybrid flow shop scheduling with parallel

batching. *International Journal of Production Economics*, 117(1):185–196, 2009. ISSN 0925-5273.

[12] E Anand, R Panneerselvam, et al. Literature review of open shop scheduling problems. *Intelligent Information Management*, 7(01):33, 2015.

[13] CJ Ancker and A Gafarian. Queueing with impatient customers who leave at random. *Journal of Industrial Engineering*, 13(84-90):171–172, 1962.

[14] A Arnaout, J Smylie, J Seely, S Robertson, K Knight, S Shin, T Ramsey, R Mallick, and J Watters. Improving breast diagnostic services with a rapid access diagnostic and support (RADS) program. *Annals of Surgical Oncology*, 20 (10):3335–3340, 2013.

[15] N Aslani and J Zhang. Integration of simulation and dea to determine the most efficient patient appointment scheduling model for a specific healthcare setting. *Journal of Industrial Engineering and Management*, 7(4):785, 2014.

[16] A Azadeh, MH Farahani, S Torabzadeh, and M Baghersad. Scheduling prioritized patients in emergency department laboratories. *Computer Methods and Programs in Biomedicine*, 117(2):61–70, 2014.

[17] A Azadeh, M Baghersad, MH Farahani, and M Zarrin. Semi-online patient scheduling in pathology laboratories. *Artificial Intelligence in Medicine*, 64(3): 217–226, 2015.

[18] M Bagheri, AG Devin, and A Izanloo. An application of stochastic programming method for nurse scheduling problem in real word hospital. *Computers & Industrial Engineering*, 96:192–200, 2016.

[19] D Bai, ZH Zhang, and Q Zhang. Flexible open shop scheduling problem to minimize makespan. *Computers & Operations Research*, 67:207–215, 2016.

[20] NTJ Bailey. A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 185–199, 1952.

[21] H Balasubramanian, S Biehl, L Dai, and A Muriel. Dynamic allocation of same-day requests in multi-physician primary care practices in the presence of prescheduled appointments. *Health Care Management Science*, 17(1):31–48, 2014.

[22] S van Balen. Hoe beïnvloedt de lay-out de looplijnen van verpleegkundigen? Een Markov-simulatie voor het UMC Utrecht. Master's thesis, University of Twente, 2015.

[23] MW Barentsz, H Wessels, PJ van Diest, RM Pijnappel, CC Van Der Pol, AJ Witkamp, MAAJ Van Den Bosch, and HM Verkooijen. Same-day diagnosis based on histology for women suspected of breast cancer: High diagnostic accuracy and favorable impact on the patient. *PloS One*, 9(7):e103105, 2014.

[24] C Barnhart, P Belobaba, and AR Odoni. Applications of operations research in the air transport industry. *Transportation Science*, 37(4):368–391, 2003.

[25] C Barz and K Rajaram. Elective patient admission and scheduling under multiple

resource constraints. *Production and Operations Management*, 24(12):1907–1930, 2015.

[26] AG Bean and J Talaga. Appointment breaking: Causes and solutions. *Marketing Health Services*, 12(4):14, 1992.

[27] J Beliën and H Forcé. Supply chain management of blood products: A literature review. *European Journal of Operational Research*, 217(1):1–16, 2012.

[28] A Bellanger and A Oulamara. Scheduling hybrid flowshop with parallel batching machines and compatibilities. *Computers and Operations Research*, 36(6):1982–1992, 2009. ISSN 0305-0548.

[29] J Benjamin-Bauman, ML Reiss, and JS Bailey. Increasing appointment keeping by reducing the call-appointment interval. *Journal of Applied Behavior Analysis*, 17(3):295–301, 1984.

[30] ML Bentaha, O Battaïa, and A Dolgui. A sample average approximation method for disassembly line balancing problem under uncertainty. *Computers & Operations Research*, 51:111–122, 2014.

[31] B Berg and BT Denton. Appointment planning and scheduling in outpatient procedure centers. In *Handbook of Healthcare System Scheduling*, pages 131–154. Springer, 2012.

[32] J van den Bergh, J Beliën, P de Bruecker, E Demeulemeester, and L de Boeck. Personnel scheduling: A literature review. *European Journal of Operational Research*, 226(3):367–385, 2013.

[33] S Bertel and JC Billaut. A genetic algorithm for an industrial multiprocessor flow shop scheduling problem with recirculation. *European Journal of Operational Research*, 159(3):651–662, 2004. ISSN 0377-2217.

[34] IA Bikker, N Kortbeek, RM van Os, and RJ Boucherie. Reducing access times for radiation treatment by aligning the doctors schemes. *Operations Research for Health Care*, 7:111–121, 2015.

[35] B Bilgin, P Demeester, M Misir, W Vancroonenburg, and G Vanden Berghe. One hyper-heuristic approach to two timetabling problems in health care. *Journal of Heuristics*, 18(3):401–434, 2012.

[36] J Bisschop and R Entriken. *AIMMS: The modeling system*. Paragon Decision Technology, 1993.

[37] EE Blæhr, R Søgaard, T Kristensen, and U Væggemose. Observational study identifies non-attendance characteristics in two hospital outpatient clinics. *Danish Medical Journal*, 63(10), 2016.

[38] N Boksmati, K Butler-Henderson, K Anderson, and T Sahama. The effectiveness

of sms reminders on appointment attendance: A meta-analysis. *Journal of Medical Systems*, 40(4):1–10, 2016.

[39] A Braaksma. *Timely and efficient planning of treatments through intelligent scheduling*, volume 15. University of Twente, 2015.

[40] A Braaksma, N Kortbeek, GF Post, and F Nollet. Integral multidisciplinary rehabilitation treatment planning. *Operations Research for Health Care*, 3(3): 145–159, 2014.

[41] A Braaksma, NM van de Vrugt, and RJ Boucherie. Online appointment scheduling: A taxonomy and review. Technical report, Department of Industrial Engineering and Management, University of Twente, 2017.

[42] S Brailsford and J Vissers. Or in healthcare: A european perspective. *European Journal of Operational Research*, 212(2):223–234, 2011.

[43] SC Brailsford, PR Harper, B Patel, and M Pitt. An analysis of the academic literature on simulation and modelling in health care. *Journal of Simulation*, 3 (3):130–140, 2009.

[44] ML Brandeau. Creating impact with operations research in health: Making room for practice in academia. *Health Care Management Science*, 19(4):305–312, 2016.

[45] L Brown. Improving histopathology turnaround time: A process management approach. *Current Diagnostic Pathology*, 10(6):444–452, 2004. ISSN 0968-6053.

[46] P Brucker, EK Burke, T Curtois, R Qu, and G Vanden Berghe. A shift sequence based approach for nurse scheduling and a new benchmark dataset. *Journal of Heuristics*, 16(4):559–573, 2010.

[47] RJ Buesa. Adapting lean to histology laboratories. *Annals of Diagnostic Pathology*, 13(5):322–33, 2009.

[48] EK Burke, P De Causmaecker, G Vanden Berghe, and H Van Landeghem. The state of the art of nurse rostering. *Journal of Scheduling*, 7(6):441–499, 2004.

[49] R Buter. Het reduceren van variatie in bedbezetting door het toewijzen van specialismen aan verpleegafdelingen. B.S. thesis, University of Twente, 2017.

[50] B Cardoen and E Demeulemeester. Capacity of clinical pathways: A strategic multi-level evaluation tool. *Journal of Medical Systems*, 32(6):443–452, 2008.

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

[52] B Cardoen, E Demeulemeester, and J Beliën. Operating room planning and scheduling: A classification scheme. *International Journal of Social Health Information Management*, 1(1):71–83, 2010.

[53] B Cardoen, E Demeulemeester, and J Beliën. Operating room planning and

scheduling: A literature review. *European Journal of Operational Research*, 201 (3):921–932, 2010.

[54] S Carpov, J Carlier, D Nace, and R Sirdey. Two-stage hybrid flow shop with precedence constraints and parallel machines at second stage. *Computers and Operations Research*, 39(3):736–745, 2012. ISSN 0305-0548.

[55] E Castro and S Petrovic. Combined mathematical programming and heuristics for a radiotherapy pre-treatment scheduling problem. *Journal of Scheduling*, 15 (3):333–346, 2012.

[56] T Cayirli and E Veral. Outpatient scheduling in health care: a review of literature. *Production and Operations Management*, 12(4):519–549, 2003.

[57] CBS. Cbs. `http://www.cbs.nl`, 2013. Online; Accessed 2017-04-06.

[58] UMC Utrecht Cancer Center. Zorgconcept, 2016.

[59] F Centorrino, MA Hernán, G Drago-Ferrante, M Rendall, A Apicella, G Längar, and RJ Baldessarini. Factors associated with noncompliance with psychiatric outpatient visits. *Psychiatric Services*, 52(3):378–380, 2001.

[60] S Ceschia and A Schaerf. Local search and lower bounds for the patient admission scheduling problem. *Computers & Operations Research*, 38(10):1452–1463, 2011.

[61] S Ceschia and A Schaerf. Dynamic patient admission scheduling with operating room constraints, flexible horizons, and patient delays. *Journal of Scheduling*, 19 (4):377–389, 2016.

[62] S Ceschia, NTT Dang, P De Causmaecker, S Haspeslagh, and A Schaerf. Second international nurse rostering competition (INRC-II)—problem description and rules—. *arXiv preprint arXiv:1501.04177*, 2015.

[63] S Chakraborty, K Muthuraman, and M Lawley. Sequential clinical scheduling with patient no-show: The impact of pre-defined slot structures. *Socio-Economic Planning Sciences*, 47(3):205–219, 2013.

[64] V Chariatte, P Michaud, A Berchtold, C Akré, and J Suris. Missed appointments in an adolescent outpatient clinic: Descriptive analyses of consultations over eight years. *Swiss Medical Weekly*, 137(47/48):677, 2007.

[65] V Chariatte, A Berchtold, C Akré, PA Michaud, and JC Suris. Missed appointments in an outpatient clinic for adolescents, an approach to predict the risk of missing. *Journal of Adolescent Health*, 43(1):38–45, 2008.

[66] M Cheng, HI Ozaku, N Kuwahara, K Kogure, and J Ota. Simulated annealing algorithm for scheduling problem in daily nursing cares. In *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on*, pages 1681–1687. IEEE, 2008.

[67] CC Chern, PS Chien, and SY Chen. A heuristic algorithm for the hospital health

examination scheduling problem. *European Journal of Operational Research*, 186 (3):1137–1157, 2008.

[68] CF Chien, YC Huang, and CH Hu. A hybrid approach of data mining and genetic algorithms for rehabilitation scheduling. *International Journal of Manufacturing Technology and Management*, 16(1-2):76–100, 2008.

[69] CF Chien, FP Tseng, and CH Chen. An evolutionary approach to rehabilitation patient scheduling: A case study. *European Journal of Operational Research*, 189 (3):1234–1253, 2008.

[70] B Cimprich. Pretreatment symptom distress in women newly diagnosed with breast cancer. *Cancer Nursing*, 22(3):185–194, 1999.

[71] Dutch Institute for Clinical Auditing (DICA). Jaarrapportage 2016, 2016.

[72] BA Clough and LM Casey. Using sms reminders in psychology clinics: A cautionary tale. *Behavioural and Cognitive Psychotherapy*, 42(03):257–268, 2014.

[73] JB Cohen, X Han, A Jemal, EM Ward, and CR Flowers. Deferred therapy is associated with improved overall survival in patients with newly diagnosed mantle cell lymphoma. *Cancer*, 122(15):2356–2363, 2016.

[74] A Condotta and NV Shakhlevich. Scheduling patient appointments via multilevel template: A case study in chemotherapy. *Operations Research for Health Care*, 3 (3):129–144, 2014.

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

[76] JP Cordier and F Riane. Towards a centralised appointments system to optimise the length of patient stay. *Decision Support Systems*, 55(2):629–639, 2013.

[77] IBM ILOG CPLEX. Ibm software group. *User-Manual CPLEX*, 12, 2011.

[78] T Curtois. Employee shift scheduling benchmark data sets. Technical report, School of Computer Science, The University of Nottingham, Nottingham, UK, 2014.

[79] J Daggy, M Lawley, D Willis, D Thayer, C Suelzer, PC DeLaurentis, A Turkcan, S Chakraborty, and L Sands. Using no-show modeling to improve clinic performance. *Health Informatics Journal*, 16(4):246–259, 2010.

[80] B Dale. *Total quality management*. Wiley Online Library, 2015.

[81] ML Davies, RM Goffman, JH May, RJ Monte, KL Rodriguez, YC Tjader, and DL Vargas. Large-scale no-show patterns and distributions for clinic operational

research. In *Healthcare*, page 15. Multidisciplinary Digital Publishing Institute, 2016.

[82] RW Day, MD Dean, R Garfinkel, and S Thompson. Improving patient flow in a hospital through dynamic allocation of cardiac diagnostic testing time slots. *Decision Support Systems*, 49(4):463–473, 2010.

[83] P Demeester, W Souffriau, P De Causmaecker, and G Vanden Berghe. A hybrid tabu search algorithm for automatically assigning patients to beds. *Artificial Intelligence in Medicine*, 48(1):61–70, 2010.

[84] E Demeulemeester, M Vanhoucke, and W Herroelen. Rangen: A random network generator for activity-on-the-node networks. *Journal of Scheduling*, 6(1):17–38, 2003.

[85] E Demirkol, S Mehta, and R Uzsoy. Benchmarks for shop scheduling problems. *European Journal of Operational Research*, 109(1):137–141, 1998.

[86] B Denton, J Viapiano, and A Vogl. Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Science*, 10(1): 13–24, 2007.

[87] BT Denton, AJ Miller, HJ Balasubramanian, and TR Huschka. Optimal allocation of surgery blocks to operating rooms under uncertainty. *Operations Research*, 58(4-part-1):802–816, 2010.

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

[89] G Dobson, S Hasija, and EJ Pinker. Reserving capacity for urgent patients in primary care. *Production and Operations Management*, 20(3):456–473, 2011.

[90] SR Downer, JG Meara, AC Da Costa, and K Sethuraman. SMS text messaging improves outpatient attendance. *Australian Health Review*, 30(3):389–396, 2006.

[91] A Drexl, R Nissen, JH Patterson, and F Salewski. Progen/πx–an instance generator for resource-constrained project scheduling problems with partially renewable resources and further extensions. *European Journal of Operational Research*, 125 (1):59–72, 2000.

[92] G Du, Z Jiang, Y Yao, and X Diao. Clinical pathways scheduling using hybrid genetic algorithm. *Journal of Medical Systems*, 37(3):9945, 2013.

[93] J Du and JYT Leung. Minimizing total tardiness on one machine is np-hard. *Mathematics of Operations Research*, 15(3):483–495, 1990.

[94] M El-Sharo, B Zheng, SW Yoon, and MT Khasawneh. An overbooking scheduling

model for outpatient appointments in a multi-provider clinic. *Operations Research for Health Care*, 6:1–10, 2015.

[95] LM Elit, EM O'Leary, GR Pond, and HY Seow. Impact of wait times on survival for women with uterine cancer. *Journal of Clinical Oncology*, 32(1):27–33, 2013.

[96] O Engin, G Ceran, and MK Yilmaz. An efficient genetic algorithm for hybrid flow shop scheduling with multiprocessor task problems. *Applied Soft Computing*, 11(3):3056–3065, 2011. ISSN 1568-4946.

[97] JT van Essen, EW Hans, JL Hurink, and A Oversberg. Minimizing the waiting time for emergency surgery. *Operations Research for Health Care*, 1(2):34–44, 2012. ISSN 2211-6923.

[98] DS Festinger, RJ Lamb, DB Marlowe, and KC Kirby. From telephone to office: Intake attendance as a function of appointment delay. *Addictive Behaviors*, 27 (1):131–137, 2002.

[99] A Fleissig, V Jenkins, S Catt, and L Fallowfield. Multidisciplinary teams in cancer care: Are they effective in the uk? *The Lancet Oncology*, 7(11):935–943, 2006.

[100] D Fone, S Hollinghurst, M Temple, A Round, N Lester, A Weightman, K Roberts, E Coyle, G Bevan, and S Palmer. Systematic review of the use and value of computer simulation modelling in population health and health care delivery. *Journal of Public Health*, 25(4):325–335, 2003.

[101] DM Foreman and M Hanna. How long can a waiting list be? *The Psychiatrist*, 24(6):211–213, 2000.

[102] PS Fournier, S Montreuil, JP Brun, C Bilodeau, and J Villa. Exploratory study to identify workload factors that have an impact on health and safety: A case study in the service sector. *Universite Laval: IRSST*, 2011.

[103] Craig M Froehle and Michael J Magazine. Improving scheduling and flow in complex outpatient clinics. In *Handbook of Healthcare Operations Management*, pages 229–250. Springer, 2013.

[104] G Gallucci, W Swartz, and F Hackerman. Brief reports: Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psychiatric Services*, 2005.

[105] Gartner. Gartner says advanced analytics is a top business priority. `http://www.gartner.com/newsroom/id/2881218`, 2014. Online; Accessed 2017-04-06.

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

[107] R Geerlings, B Aldenkamp, L Gottmer-Welschen, P de With, S Zinger, A van Staa, and A de Louw. Evaluation of a multidisciplinary epilepsy transition clinic for adolescents. *European Journal of Paediatric Neurology*, 2016.

[108] S van der Geer, M Frunt, H Romero, N Dellaert, M Jansen-Vullers, T Demeyere, M Neumann, and G Krekels. One-stop-shop treatment for basal cell carcinoma,

part of a new disease management strategy. *Journal of the European Academy of Dermatology and Venereology*, 26(9):1154–1157, 2012.

[109] H Gehring and J Homberger. A parallel two-phase metaheuristic for routing problems with time windows. *Asia-Pacific Journal of Operational Research*, 18 (1):35, 2001.

[110] ZIM Geuke and G Sturtz. Patiëntenvoorkeuren in de diagnostiek van kanker. B.Sc. thesis, University of Twente, 2016.

[111] E Ghafari and R Sahraeian. A two-stage hybrid flowshop scheduling problem with serial batching. *International Journal of Industrial Engineering and Production Research*, 25(1):55–63, 2014.

[112] EM Goldratt. *What is this thing called theory of constraints and how should it be implemented?* North River Press, 1990.

[113] A Goodridge, D Woodhouse, and J Barbour. Improving patient access at a movement disorder clinic by participating in a process improvement program. *BMJ Quality Improvement Reports*, 2(1):u479–w1007, 2013.

[114] RL Graham, EL Lawler, JK Lenstra, and AHGR Kan. Optimization and approximation in deterministic sequencing and scheduling: A survey. *Annals of Discrete Mathematics*, 5:287–326, 1979.

[115] RE Gray, MI Fitch, C Phillips, M Labrecque, and L Klotz. Presurgery experiences of prostate cancer patients and their spouses. *Cancer Practice*, 7(3):130–135, 1999.

[116] LV Green and S Savin. Reducing delays for medical appointments: A queueing approach. *Operations Research*, 56(6):1526–1538, 2008.

[117] JD Griffiths, JE Williams, and RM Wood. Scheduling physiotherapy treatment in an inpatient setting. *Operations Research for Health Care*, 1(4):65–72, 2012.

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

[119] D Gupta and B Denton. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, 40(9):800–819, 2008.

[120] JND Gupta. Two-stage, hybrid flowshop scheduling problem. *Journal of the Operational Research Society*, pages 359–364, 1988. ISSN 0160-5682.

[121] JND Gupta, AMA Hariri, and CN Potts. Scheduling a two-stage hybrid flow shop with parallel machines at the first stage. *Annals of Operations Research*, 69: 171–191, 1997. ISSN 0254-5330.

[122] S Gupta and IA Karimi. An improved milp formulation for scheduling multiproduct, multistage batch plants. *Industrial and Engineering Chemistry Research*, 42 (11):2365–2380, 2003. ISSN 0888-5885.

[123] CE Guse, L Richardson, M Carle, and K Schmidt. The effect of exit-interview

patient education on no-show rates at a family practice residency clinic. *The Journal of the American Board of Family Practice*, 16(5):399–404, 2003.

[124] R Guy, J Hocking, H Wand, S Stott, H Ali, and J Kaldor. How effective are short message service reminders at increasing clinic attendance? A meta-analysis and systematic review. *Health Services Research*, 47(2):614–632, 2012.

[125] S Hahn-Goldberg, MW Carter, JC Beck, M Trudeau, P Sousa, and K Beattie. Dynamic optimization of chemotherapy outpatient scheduling with uncertainty. *Health Care Management Science*, 17(4):379–392, 2014.

[126] W Hamilton, A Round, and D Sharp. Patient, hospital, and general practitioner characteristics associated with non-attendance: A cohort study. *British Journal of General Practice*, 52(477):317–319, 2002.

[127] EW Hans, W Herroelen, R Leus, and G Wullink. A hierarchical approach to multi-project planning under uncertainty. *Omega*, 35(5):563–577, 2007.

[128] EW Hans, G Wullink, M Van Houdenhoven, and G Kazemier. Robust surgery loading. *European Journal of Operational Research*, 185(3):1038–1050, 2008.

[129] EW Hans, M Van Houdenhoven, and PJH Hulshof. A framework for healthcare planning and control. In *Handbook of healthcare system scheduling*, pages 303–320. Springer, 2012.

[130] I Harjunkoski and IE Grossmann. Decomposition techniques for multistage scheduling problems using mixed-integer and constraint programming methods. *Computers and Chemical Engineering*, 26(11):1533–1552, 2002. ISSN 0098-1354.

[131] I Harjunkoski, CT Maravelias, P Bongers, PM Castro, S Engell, IE Grossmann, J Hooker, C Mndez, G Sand, and J Wassick. Scope for industrial applications of production scheduling models and solution methods. *Computers and Chemical Engineering*, 62:161–193, 2014. ISSN 0098-1354.

[132] S Harris. *Essays in appointment management*. PhD thesis, University of Pittsburgh, 2016.

[133] MC van Harten, FJP Hoebers, KW Kross, ED van Werkhoven, MWM van den Brekel, and BAC van Dijk. Determinants of treatment waiting times for head and neck cancer in the Netherlands and their relation to survival. *Oral Oncology*, 51(3):272–278, 2015.

[134] WH van Harten, EW Hans, and WAM van Lent. Aanpak efficiency te ondoordacht. *Medisch Contact*, 65(6):264, 2010.

[135] S Haspeslagh, P De Causmaecker, A Schaerf, and M Stølevik. The first interna-

tional nurse rostering competition 2010. *Annals of Operations Research*, 218(1): 221–236, 2014.

[136] DSJ Hawker. Increasing initial attendance at mental health out-patient clinics: opt-in systems and other interventions. *The Psychiatrist*, 31(5):179–182, 2007.

[137] JMG Heerkens and A van Winden. *Solving managerial problems systematically*. Noordhoff Uitgevers, 2017.

[138] WJP Heijnen. Optimizing OR scheduling for the ophthalmology department. B.S. thesis, University of Twente, 2016.

[139] Y Hellstadius, J Lagergren, J Zylstra, J Gossage, A Davies, CM Hultman, P Lagergren, and Anna Wikman. Prevalence and predictors of anxiety and depression among esophageal cancer patients prior to surgery. *Diseases of the Esophagus*, 30(8):1–7, 2017.

[140] R Henderson. Encouraging attendance at outpatient appointments: Can we do more? *Scottish Medical Journal*, 53(1):9–12, 2008.

[141] J Homberger. A (μ, λ)-coordination mechanism for agent-based multi-project scheduling. *OR Spectrum*, 34(1):107–132, 2012.

[142] JA Hoogeveen, JK Lenstra, and B Veltman. Preemptive scheduling in a two-stage multiprocessor flow shop is np-hard. *European Journal of Operational Research*, 89(1):172–175, 1996. ISSN 0377-2217.

[143] X Hu, H Wu, S Zhang, X Dai, and Y Jin. Scheduling outpatients in hospital examination departments. In *Industrial Engineering and Engineering Management, 2009. IEEM 2009. IEEE International Conference on*, pages 335–338. IEEE, 2009.

[144] Y Huang and P Zuniga. Dynamic overbooking scheduling system to improve patient access. *Journal of the Operational Research Society*, 63(6):810–820, 2012.

[145] P Hulshof, N Kortbeek, RJ Boucherie, EW Hans, and P Bakker. Taxonomic classification of planning decisions in health care: A structured review of the state of the art in OR/MS. *Health Systems*, 1(2):129–175, 2012.

[146] PJH Hulshof, RJ Boucherie, EW Hans, and JL Hurink. Tactical resource allocation and elective patient admission planning in care processes. *Health Care Management Science*, 16(2):152–166, 2013.

[147] PJH Hulshof, MRK Mes, RJ Boucherie, and EW Hans. Patient admission planning using approximate dynamic programming. *Flexible Services and Manufacturing Journal*, 28(1-2):30–61, 2016.

[148] SV Jerić and JR Figueira. Multi-objective scheduling and a resource allocation problem in hospitals. *Journal of Scheduling*, 15(5):513–535, 2012.

[149] HR Jocham, T Dassen, G Widdershoven, and R Halfens. Quality of life in pal-

liative care cancer patients: A literature review. *Journal of Clinical Nursing*, 15 (9):1188–1195, 2006.

[150] SM Johnson. Optimal two- and three-stage production schedules with setup times included. *Naval Research Logistics Quarterly*, 1(1):61–68, 1954. ISSN 1931-9193.

[151] RV Jones and B Greenwood. Breast cancer: causes of patients' distress identified by qualitative analysis. *British Journal of General Practice*, 44(385):370–371, 1994.

[152] JB Jun, SH Jacobson, and JR Swisher. Application of discrete-event simulation in health care clinics: A survey. *Journal of the Operational Research Society*, pages 109–123, 1999.

[153] AG Kalton, MR Singh, DA August, CM Parin, and EJ Othman. Using simulation to improve the operational efficiency of a multidisciplinary clinic. *Journal of the Society for Health Systems*, 5(3):43–62, 1997.

[154] KWF Kankerbestrijding. Advies inzake wachttijdnormen in de kankerzorg, 2006.

[155] AS Kapadia, SE Vineberg, and CD Rossi. Predicting course of treatment in a rehabilitation hospital: a markovian model. *Computers & Operations Research*, 12(5):459–469, 1985.

[156] T Kapamara, K Sheibani, D Petrovic, O Haas, and C Reeves. A simulation of a radiotherapy treatment system: A case study of a local cancer centre. In *ORP3 Meeting*, 2007.

[157] YD Kim. Heuristics for flowshop scheduling problems minimizing mean tardiness. *Journal of the Operational Research Society*, pages 19–28, 1993.

[158] K Klassen and T Rohleder. Scheduling outpatient appointments in a dynamic environment. *Journal of Operations Management*, 14(2):83–101, 1996.

[159] A Kleywegt, A Shapiro, and T Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

[160] AL Kok, CM Meyer, H Kopfer, and JMJ Schutten. A dynamic programming heuristic for the vehicle routing problem with time windows and european community social legislation. *Transportation Science*, 44(4):442–454, 2010.

[161] R Kolisch and A Sprecher. Psplib-a project scheduling problem library: Or software-orsep operations research software exchange program. *European Journal of Operational Research*, 96(1):205–216, 1997.

[162] R Kolisch, A Sprecher, and A Drexl. Characterization and generation of a general class of resource-constrained project scheduling problems. *Management science*, 41(10):1693–1703, 1995.

[163] R Kolisch, C Schwindt, and A Sprecher. Benchmark instances for project scheduling problems. In *Project Scheduling*, pages 197–212. Springer, 1999.

[164] N Kortbeek, ME Zonderland, A Braaksma, IMH Vliegen, RJ Boucherie, N Litvak, and EW Hans. Designing cyclic appointment schedules for outpatient clinics with

scheduled and unscheduled patient arrivals. *Performance Evaluation*, 80:5–26, 2014.

[165] E Koshy, J Car, and A Majeed. Effectiveness of mobile-phone short message service (sms) reminders for ophthalmology outpatient appointments: Observational study. *BMC Ophthalmology*, 8(1):1, 2008.

[166] G Kunigiri, N Gajebasia, and D Sallah. Improving attendance in psychiatric outpatient clinics by using reminders. *Journal of Telemedicine and Telecare*, page 1357633X14555642, 2014.

[167] LR LaGanga and SR Lawrence. Clinic overbooking to improve patient access and increase provider productivity. *Decision Sciences*, 38(2):251–276, 2007.

[168] LR LaGanga and SR Lawrence. Appointment overbooking in health care clinics to improve patient service and clinic performance. *Production and Operations Management*, 21(5):874–888, 2012.

[169] G Lamé, O Jouini, and J Stal-Le Cardinal. Outpatient chemotherapy planning: A literature review with insights from a case study. *IIE Transactions on Healthcare Systems Engineering*, 6(3):127–139, 2016.

[170] M Lamiri, F Grimaud, and X Xie. Optimization methods for a stochastic surgery planning problem. *International Journal of Production Economics*, 120(2):400–410, 2009.

[171] AM Law. *Simulation modeling and analysis*. McGraw-Hill, 2007.

[172] R Lebcir, E Demir, R Ahmad, C Vasilakis, and D Southern. A discrete event simulation model to evaluate the use of community services in the treatment of patients with parkinsons disease in the united kingdom. *BMC Health Services Research*, 17(1):50, 2017.

[173] VJ Lee, A Earnest, MI Chen, and B Krishnan. Predictors of failed attendances in a multi-specialty outpatient centre using electronic databases. *BMC Health Services Research*, 5(1):1, 2005.

[174] AG Leeftink and EW Hans. Case mix classification and a benchmark set for surgery scheduling. *Journal of Scheduling*, pages 1–17, 2017.

[175] AG Leeftink, RJ Boucherie, EW Hans, MAM Verdaasdonk, IMH Vliegen, and PJ van Diest. Batch scheduling in the histopathology laboratory. *Flexible Services and Manufacturing Journal*, 2016. doi: 10.1007/s10696-016-9257-3.

[176] AG Leeftink, RJ Boucherie, EW Hans, MAM Verdaasdonk, IMH Vliegen, and PJ van Diest. Predicting turnaround time reductions of the diagnostic track in the histopathology laboratory using mathematical modelling. *Journal of Clinical Pathology*, pages jclinpath–2015, 2016.

[177] AG Leeftink, IMH Vliegen, and EW Hans. Stochastic integer programming for multi-disciplinary clinic planning. *Health Care Management Science*, 2017.

[178] B Liang, A Turkcan, ME Ceyhan, and K Stuart. Improvement of chemotherapy

patient flow and scheduling in an outpatient oncology clinic. *International Journal of Production Research*, 53(24):7177–7190, 2015.

[179] MN Liao, MF Chen, SC Chen, and PL Chen. Uncertainty and anxiety during the diagnostic period for women with suspected breast cancer. *Cancer Nursing*, 31(4):274–283, 2008.

[180] LS Lim and P Varkey. E-mail reminders: a novel method to reduce outpatient clinic nonattendance. *The Internet Journal of Healthcare Administration*, 3(1), 2005.

[181] CKY Lin. An adaptive scheduling heuristic with memory for the block appointment system of an outpatient specialty clinic. *International Journal of Production Research*, 53(24):7488–7516, 2015.

[182] G Litton, D Kane, G Clay, P Kruger, T Belnap, and B Parkinson. Multidisciplinary cancer care with a patient and physician satisfaction focus. *Journal of Oncology Practice*, 6(6):e35–e37, 2010.

[183] B Liu, L Wang, Y Liu, B Qian, and YH Jin. An effective hybrid particle swarm optimization for batch scheduling of polypropylene processes. *Computers and Chemical Engineering*, 34(4):518–528, 2010. ISSN 0098-1354.

[184] N Liu. Optimal choice for appointment scheduling window under patient no-show behavior. *Production and Operations Management*, 25(1):128–142, 2016.

[185] N Liu and S Ziya. Panel size and overbooking decisions for appointment-based services under patient no-shows. *Production and Operations Management*, 23(12): 2209–2223, 2014.

[186] WS Lovejoy and Y Li. Hospital operating room capacity expansion. *Management Science*, 48(11):1369–1387, 2002.

[187] H Luo, GQ Huang, Y Feng Zhang, and Q Yun Dai. Hybrid flowshop scheduling with batch-discrete processors and machine maintenance in time windows. *International Journal of Production Research*, 49(6):1575–1603, 2011. ISSN 0020-7543.

[188] X Ma, A Sauré, M Puterman, M Taylor, and S Tyldesley. Capacity planning and appointment scheduling for new patient oncology consults. *Health Care Management Science*, pages 1–15, 2015.

[189] C Mannino, EJ Nilssen, and TE Nordlander. Sintef ict: Mss-adjusts surgery data. `https://www.sintef.no/Projectweb/Health-care-optimization/Testbed/`, 2010.

[190] E Marcon, S Kharraja, and G Simonnet. The operating theatre planning by the follow-up of the risk of no realization. *International Journal of Production Economics*, 85(1):83–90, 2003.

[191] RE Mark, PL Klarenbeek, GJM Rutten, and MM Sitskoorn. Why dont neurosurgery patients return for neuropsychological follow-up? predictors for voluntary

appointment keeping and reasons for cancellation. *The Clinical Neuropsychologist*, 28(1):49–64, 2014.

[192] JT Markowitz, LK Volkening, and LMB Laffel. Care utilization in a pediatric diabetes clinic: cancellations, parental attendance, and mental health appointments. *The Journal of Pediatrics*, 164(6):1384–1389, 2014.

[193] I Marques, ME Captivo, and MV Pato. An integer programming approach to elective surgery scheduling. *OR Spectrum*, 34(2):407–427, 2012.

[194] M Marshall, C Pagel, C French, M Utley, D Allwood, N Fulop, C Pope, V Banks, and A Goldmann. Moving improvement research closer to practice: the researcher-in-residence model. *BMJ Quality & Safety*, 23(10):801–805, 2014.

[195] J Marynissen and E Demeulemeester. Literature review on integrated hospital scheduling problems. Technical report, Faculty of Economics and Business, KU Leuven, 2016.

[196] M Matta and S Patterson. Evaluating multiple performance measures across several dimensions at a multi-facility outpatient center. *Health Care Management Science*, 10(2):173–194, 2007.

[197] ME Matta. A genetic algorithm for the proportionate multiprocessor open shop. *Computers & Operations Research*, 36(9):2601–2618, 2009.

[198] ME Matta and SE Elmaghraby. Polynomial time algorithms for two special classes of the proportionate multiprocessor open shop. *European Journal of Operational Research*, 201(3):720–728, 2010.

[199] JH May, DP Strum, and LG Vargas. Fitting the lognormal distribution to surgical procedure times. *Decision Sciences*, 31(1):129–148, 2000.

[200] JM McLaughlin, RT Anderson, AK Ferketich, EE Seiber, R Balkrishnan, and ED Paskett. Effect on survival of longer intervals between confirmed diagnosis and treatment initiation among low-income women with breast cancer. *Journal of Clinical Oncology*, 30(36):4493–4500, 2012.

[201] SR McLean, D Karsanji, J Wilson, E Dixon, FR Sutherland, J Pasieka, C Ball, and OF Bathe. The effect of wait times on oncological outcomes from periampullary adenocarcinomas. *Journal of Surgical Oncology*, 107(8):853–858, 2013.

[202] TF Meijman, G Mulder, PJD Drenth, and H Thierry. *Psychological aspects of workload*, volume 2, pages 5–33. Psychology Press, Hove, U.K., 1998.

[203] CA Mendez, J Cerda, IE Grossmann, I Harjunkoski, and M Fahl. State-of-the-art review of optimization methods for short-term scheduling of batch processes. *Computers and Chemical Engineering*, 30(6):913–946, 2006. ISSN 0098-1354.

[204] Rijksinstituut voor Volksgezondheid en Milieu (RIVM). Een samenhangend beeld van kanker: Ziekte, zorg, mens en maatschappij, 2016.

[205] D Min and Y Yih. Scheduling elective surgery under uncertainty and downstream

capacity constraints. *European Journal of Operational Research*, 206(3):642–652, 2010.

[206] HS Mirsanei, M Zandieh, MJ Moayed, and MR Khabbazi. A simulated annealing algorithm approach to hybrid flow shop scheduling with sequence-dependent setup times. *Journal of Intelligent Manufacturing*, 22(6):965–978, 2011. ISSN 0956-5515.

[207] JM Molina-Pariente, EW Hans, JM Framinan, and T Gomez-Cia. New heuristics for planning operating rooms. *Computers & Industrial Engineering*, 90:429–443, 2015.

[208] CG Moore, P Wilson-Witherspoon, and JC Probst. Time and money: effects of no-shows at a family practice residency clinic. *Family Medicine*, 33(7):522–527, 2001.

[209] V Mor, S Allen, and M Malin. The psychosocial impact of cancer on older versus younger patients and their families. *Cancer*, 74(7):2118–2127, 1994.

[210] A Morales, H Essenfeld, E Essenfeld, MC Duboue, V Vincek, and M Nadji. Continuous-specimen-flow, high-throughput, 1-hour tissue processing: a system for rapid diagnostic tissue preparation. *Archives of Pathology and Laboratory Medicine*, 126(5):583–590, 2002.

[211] L Mruşter, T Weijters, G de Vries, A van den Bosch, and W Daelemans. Logistic-based patient grouping for multi-disciplinary treatment. *Artificial Intelligence in Medicine*, 26(1):87–107, 2002.

[212] D Muirhead, P Aoun, M Powell, F Juncker, and J Mollerup. Pathology economic model tool: A novel approach to workflow and budget cost analysis in an anatomic pathology laboratory. *Archives of Pathology and Laboratory Medicine*, 134(8): 1164–1169, 2010.

[213] J Munkholm, ML Talman, and T Hasselager. Implementation of a new rapid tissue processing method–advantages and challenges. *Pathology-Research and Practice*, 204(12):899–904, 2008.

[214] P Murchie, EA Raja, DH Brewster, NC Campbell, LD Ritchie, R Robertson, L Samuel, N Gray, and AJ Lee. Time from first presentation in primary care to treatment of symptomatic colorectal cancer: Effect on disease stage and survival. *British Journal of Cancer*, 111(3):461–469, 2014.

[215] M Murray and C Tantau. Redefining open access to primary care. *Managed Care Quarterly*, 7:45–55, 1999.

[216] N Musliu, A Schaerf, and W Slany. Local search for shift design. *European Journal of Operational Research*, 153(1):51–64, 2004.

[217] N Mustafee, K Katsaliaki, and SJE Taylor. Profiling literature in healthcare simulation. *Simulation*, 86(8-9):543–558, 2010.

[218] S Mutlu, J Benneyan, J Terrell, V Jordan, and A Turkcan. A co-availability

scheduling model for coordinating multi-disciplinary care teams. *International Journal of Production Research*, 53(24):7226–7237, 2015.

[219] RD Neal, P Tharmanathan, B France, NU Din, S Cotton, J Fallon-Ferguson, W Hamilton, A Hendry, M Hendry, R Lewis, et al. Is increased time to diagnosis and treatment in symptomatic cancer associated with poorer outcomes? Systematic review. *British Journal of Cancer*, 112:S92–S107, 2015.

[220] DL Nguyen, RS DeJesus, and ML Wieland. Missed appointments in resident continuity clinic: Patient characteristics and health care outcomes. *Journal of Graduate Medical Education*, 3(3):350–355, 2011.

[221] JB Norris, C Kumar, S Chand, H Moskowitz, SA Shade, and DR Willis. An empirical investigation into factors affecting patient cancellations and no-shows at outpatient clinics. *Decision Support Systems*, 57:428–443, 2014.

[222] HJ Oh, A Muriel, H Balasubramanian, K Atkinson, and T Ptaszkiewicz. Guidelines for scheduling in primary care under different patient types and stochastic nurse and provider service times. *IIE Transactions on Healthcare Systems Engineering*, 3(4):263–279, 2013.

[223] HJA Oh, A Muriel, and H Balasubramanian. A user-friendly excel simulation for scheduling in primary care practices. In *Simulation Conference (WSC), 2014 Winter*, pages 1177–1185. IEEE, 2014.

[224] Benchmarking OK. Benchmarking OK, leren van elkaar. `http://benchmarking-ok.nl`, 2017. Accessed: 2017-05-17.

[225] Taskforce Oncologie. Koersboek oncologische netwerkvorming, 2015.

[226] WHWM Otten. Pooling hospital beds: A capacity allocation study within the Wilhelmina Kinderziekenhuis. B.S. thesis, University of Twente, 2017.

[227] A Parikh, K Gupta, AC Wilson, K Fields, NM Cosgrove, and JB Kostis. The effectiveness of outpatient appointment reminder systems in reducing no-show rates. *The American Journal of Medicine*, 123(6):542–548, 2010.

[228] MR Partin, A Gravely, ZF Gellad, S Nugent, JF Burgess, A Shaukat, and DB Nelson. Factors associated with missed and cancelled colonoscopy appointments at veterans health administration facilities. *Clinical Gastroenterology and Hepatology*, 14(2):259–267, 2016.

[229] MI Patel, DT DeConcini, E Lopez-Corona, M Ohori, T Wheeler, and PT Scardino. An analysis of men with clinically localized prostate cancer who deferred definitive therapy. *The Journal of Urology*, 171(4):1520–1524, 2004.

[230] S Patel, JB Smith, E Kurbatova, and J Guarner. Factors that impact turnaround time of surgical pathology specimens in an academic institution. *Human Pathology*, 43(9):1501–1505, 2012.

[231] C Paul, M Carey, A Anderson, L Mackenzie, R Sanson-Fisher, R Courtney, and T Clinton-McHarg. Cancer patients' concerns regarding access to cancer care:

Perceived impact of waiting times along the diagnosis and treatment journey. *European Journal of Cancer Care*, 21(3):321–329, 2012.

[232] TO Paulussen, NR Jennings, KS Decker, and A Heinzl. Distributed patient scheduling in hospitals. In *International Joint Conference of Artificial Intelligence (IJCAI)*, pages 1224–1229, 2003.

[233] Y Peng, X Qu, and J Shi. A hybrid simulation and genetic algorithm approach to determine the optimal scheduling templates for open access clinics admitting walk-in patients. *Computers & Industrial Engineering*, 72:282–296, 2014.

[234] Y Peng, E Erdem, J Shi, C Masek, and P Woodbridge. Large-scale assessment of missed opportunity risks in a complex hospital setting. *Informatics for Health and Social Care*, 41(2):112–127, 2016.

[235] E Pérez, L Ntaimo, WE Wilhelm, C Bailey, and P McCormack. Patient and resource scheduling of multi-step medical procedures in nuclear medicine. *IIE Transactions on Healthcare Systems Engineering*, 1(3):168–184, 2011.

[236] E Pérez, L Ntaimo, CO Malavé, C Bailey, and P McCormack. Stochastic online appointment scheduling of multi-step sequential procedures in nuclear medicine. *Health Care Management Science*, 16(4):281–299, 2013.

[237] NJ Perron, MD Dao, NC Righini, JP Humair, B Broers, F Narring, DM Haller, and JM Gaspoz. Text-messaging versus telephone reminders to reduce missed appointments in an academic primary care clinic: A randomized controlled trial. *BMC Health Services Research*, 13(1):1, 2013.

[238] D Petrovic, M Morshed, and S Petrovic. Multi-objective genetic algorithms for scheduling of radiotherapy treatments for categorised cancer patients. *Expert Systems with Applications*, 38(6):6994–7002, 2011.

[239] D Petrovic, E Castro, S Petrovic, and T Kapamara. Radiotherapy scheduling. In *Automated Scheduling and Planning*, pages 155–189. Springer, 2013.

[240] V Pillac, C Gueret, and AL Medaglia. A parallel matheuristic for the technician routing and scheduling problem. *Optimization Letters*, 7(7):1525–1535, 2013.

[241] V Podgorelec and P Kokol. Genetic algorithm based system for patient scheduling in highly constrained situations. *Journal of Medical Systems*, 21(6):417–427, 1997.

[242] ME Porter. What is value in health care? *New England Journal of Medicine*, 363 (26):2477–2481, 2010.

[243] CN Potts and MY Kovalyov. Scheduling with batching: A review. *European Journal of Operational Research*, 120(2):228–249, 2000. ISSN 0377-2217.

[244] P Prasad and CT Maravelias. Batch selection, assignment and sequencing in

multi-stage multi-product processes. *Computers and Chemical Engineering*, 32 (6):1106–1119, 2008. ISSN 0098-1354.

[245] S Proctor, B Lehaney, C Reeves, and Z Khan. Modelling patient flow in a radiotherapy department. *OR Insight*, 20(3):6–14, 2007.

[246] T Pyzdek and PA Keller. *The six sigma handbook*. McGraw-Hill Education New York, 2014.

[247] R Qu, EK Burke, B McCollum, LTG Merlot, and SY Lee. A survey of search methodologies and automated system development for examination timetabling. *Journal of Scheduling*, 12(1):55–89, 2009.

[248] X Qu, R Rardin, JA Williams, and D Willis. Matching daily healthcare provider capacity to demand in advanced access scheduling systems. *European Journal of Operational Research*, 183(2):812–826, 2007.

[249] X Qu, R Rardin, and JA Williams. A mean–variance model to optimize the fixed versus open appointment percentages in open access scheduling systems. *Decision Support Systems*, 53(3):554–564, 2012.

[250] X Qu, Y Peng, N Kong, and J Shi. A two-phase approach to scheduling multicategory outpatient appointments–a case study of a womens clinic. *Health Care Management Science*, 16(3):197–216, 2013.

[251] X Qu, Y Peng, J Shi, and L LaGanga. An mdp model for walk-in patient admission management in primary care clinics. *International Journal of Production Economics*, 168:303–320, 2015.

[252] M Ramos, M Esteva, E Cabeza, C Campillo, J Llobera, and A Aguiló. Relationship of diagnostic and therapeutic delay with survival in colorectal cancer: A review. *European Journal of Cancer*, 43(17):2467–2478, 2007.

[253] Y Rapoport, S Kreitler, S Chaitchik, R Algor, and K Weissler. Psychosocial problems in head-and-neck cancer patients and their change with time since diagnosis. *Annals of Oncology*, 4(1):69–73, 1993.

[254] IM Raschendorfer and HW Hamacher. *Hierarchical edge colorings and rehabilitation therapy planning in Germany*. Technische Universität Kaiserslautern, Fachbereich Mathematik, 2014.

[255] A Ratcliffe, W Gilland, and A Marucheck. Revenue management for outpatient appointments: Joint capacity control and overbooking with class-dependent noshows. *Flexible Services and Manufacturing Journal*, 24(4):516–548, 2012.

[256] I Ribas, R Leisten, and JM Framinan. Review and classification of hybrid flow shop scheduling problems from a production system and a solutions procedure perspective. *Computers and Operations Research*, 37(8):1439–1454, 2010. ISSN 0305-0548.

[257] T Ribé, T Ribalta, R Lledó, G Torras, MA Asenjo, and A Cardesa. Evaluation

of turnaround times as a component of quality assurance in surgical pathology. *International Journal for Quality in Health Care*, 10(3):241–245, 1998.

[258] MA Richards, AM Westcombe, SB Love, P Littlejohns, and AJ Ramirez. Influence of delay on survival in patients with breast cancer: A systematic review. *The Lancet*, 353(9159):1119–1126, 1999.

[259] C van Riet and E Demeulemeester. Trade-offs in operating room planning for electives and emergencies: A review. *Operations Research for Health Care*, 7: 52–69, 2015.

[260] A Riise and EK Burke. Local search for the surgery admission planning problem. *Journal of Heuristics*, 17(4):389–414, 2011.

[261] LW Robinson and RR Chen. A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing & Service Operations Management*, 12(2):330–346, 2010.

[262] H Romero, N Dellaert, S van der Geer, M Frunt, M Jansen-Vullers, and G Krekels. Admission and capacity planning for the implementation of one-stop-shop in skin cancer treatment using simulation-based optimization. *Health Care Management Science*, 16(1):75–86, 2013.

[263] A Rossi, A Puppato, and M Lanzetta. Heuristics for scheduling a two-stage hybrid flow shop with parallel batching machines: application at a hospital sterilisation plant. *International Journal of Production Research*, 51(8):2363–2376, 2013. ISSN 0020-7543.

[264] R Ruiz and JA Vazquez-Rodriguez. The hybrid flow shop scheduling problem. *European Journal of Operational Research*, 205(1):1–18, 2010. ISSN 0377-2217.

[265] NEH Saadani, Z Bahroun, and A Bouras. A linear mathematical model for patients' activities scheduling on hospital resources. In *Control, Decision and Information Technologies (CoDIT), 2014 International Conference on*, pages 074–080. IEEE, 2014.

[266] A Sadki, X Xie, and F Chauvin. Appointment scheduling of oncology outpatients. In *Automation Science and Engineering (CASE), 2011 IEEE Conference on*, pages 513–518. IEEE, 2011.

[267] C Samuel, K Gonapa, PK Chaudhary, and A Mishra. Supply chain dynamics in healthcare services. *International Journal of Health Care Quality and Assurance*, 23:631–642, 2010.

[268] P Santibáñez, V Chow, J French, M Puterman, and S Tyldesley. Reducing patient wait times and improving resource utilization at British Columbia Cancer Agencys ambulatory care unit through simulation. *Health Care Management Science*, 12 (4):392–407, 2009.

[269] A Saremi, P Jula, T ElMekkawy, and GG Wang. Bi-criteria appointment scheduling of patients with heterogeneous service sequences. *Expert Systems with Applications*, 42(8):4029–4041, 2015.

[270] E van Sark. Care on demand-improving the coordination between the members

of the multidisciplinary team of endocrine oncology. Master's thesis, University of Twente, 2016.

[271] WB Schaufeli and AB Bakker. Job demands, job resources, and their relationship with burnout and engagement: A multisample study. *Journal of Organizational Behavior*, 25(3):293–315, 2004. ISSN 1099-1379.

[272] O Schauman, LE Aschan, N Arias, S Beards, and S Clement. Interventions to increase initial appointment attendance in mental health services: A systematic review. *Psychiatric Services*, 64(12):1249–1258, 2013.

[273] K Schimmelpfeng, S Helber, and S Kasper. Decision support for rehabilitation hospital scheduling. *OR spectrum*, 34(2):461–489, 2012.

[274] L Serrano, P Hegge, B Sato, B Richmond, and L Stahnke. Using lean principles to improve quality, patient safety, and workflow in histology and anatomic pathology. *Advances in Anatomic Pathology*, 17(3):215–221, 2010.

[275] MJ Sewitch and M Hosseina. Cancelled and missed colonoscopy appointments not easy to measure. *Clinical Gastroenterology and Hepatology*, 14(3):485–486, 2016.

[276] SJ Shah, P Cronin, CS Hong, AS Hwang, JM Ashburner, BI Bearnot, CA Richardson, BW Fosburgh, and AB Kimball. Targeted reminder phone calls to patients at high risk of no-show for primary care appointment: A randomized trial. *Journal of General Internal Medicine*, pages 1–7, 2016.

[277] SH Shin, BJ Choi, SM Ryew, JW Kim, DS Kim, WK Chung, HR Choi, and JC Koo. Development of an improved scheduling algorithm for lab test operations on a small-size bio robot platform. *JALA: Journal of the Association for Laboratory Automation*, 15(1):15–24, 2010.

[278] P Skrabek. Importance of accessible cancer care. *Transfusion and Apheresis Science*, 49(2):139–143, 2013.

[279] N Slack, S Chambers, and R Johnston. *Operations management*. Pearson Education, 2010.

[280] MM Solomon. Algorithms for the vehicle routing and scheduling problems with time window constraints. *Operations Research*, 35(2):254–265, 1987.

[281] MHP Sommers. Emergencies in planning and planning emergencies: Research to the operating room planning for emergency patients at UMC Utrecht. Master's thesis, University of Twente, 2016.

[282] Stichting Oncologische Samenwerking (SONCOS). Multidisciplinaire normering oncologische zorg in nederland, 2017.

[283] JR Sorensen, J Johansen, L Gano, JA Sørensen, SR Larsen, PB Andersen, A Thomassen, and C Godballe. A package solution fast track program can re-

duce the diagnostic waiting time in head and neck cancer. *European Archives of Oto-Rhino-Laryngology*, 271(5):1163–1170, 2014.

[284] TJ Speldekamp. Inzicht in de bedbezetting en personele inzet binnen een verpleegafdeling. B.S. thesis, University of Twente, 2016.

[285] PS Stepaniak, C Heij, GHH Mannaerts, M de Quelerij, and G de Vries. Modeling procedure and surgical times for current procedural terminology-anesthesia-surgeon combinations and evaluation in terms of case-duration prediction and operating room efficiency: a multicenter study. *Anesthesia & Analgesia*, 109(4): 1232–1245, 2009.

[286] BA Stotler and A Kratz. Determination of turnaround time in the clinical laboratory. *American Journal of Clinical Pathology*, 138(5):724–729, 2012.

[287] R Sullivan, J Peppercorn, K Sikora, J Zalcberg, NJ Meropol, E Amir, D Khayat, P Boyle, P Autier, IF Tannock, et al. Delivering affordable cancer care in high-income countries. *The Lancet Oncology*, 12(10):933–980, 2011.

[288] MJ Taylor, C McNicholas, C Nicolay, A Darzi, D Bell, and JE Reed. Systematic review of the application of the plan–do–study–act method to improve quality in healthcare. *BMJ Qual Saf*, 23(4):290–298, 2014.

[289] NF Taylor, J Bottrell, K Lawler, and D Benjamin. Mobile telephone short message service reminders can reduce nonattendance in physical therapy outpatient clinics: a randomized controlled trial. *Archives of Physical Medicine and Rehabilitation*, 93(1):21–26, 2012.

[290] VA Truong. Optimal advance scheduling. *Management Science*, 61(7):1584–1597, 2015.

[291] PFJ Tsai and GY Teng. A stochastic appointment scheduling system on multiple resources with dynamic call-in sequence and patient no-shows for an outpatient clinic. *European Journal of Operational Research*, 239(2):427–436, 2014.

[292] H Tsubone, M Ohba, and T Uetake. The impact of lot sizing and sequencing on manufacturing performance in a two-stage hybrid flow shop. *International Journal of Production Research*, 34(11):3037–3053, 1996. ISSN 0020-7543.

[293] DC Tyler, CA Pasquariello, and CH Chen. Determining optimum operating room utilization. *Anesthesia & Analgesia*, 96(4):1114–1121, 2003.

[294] Universitair Medisch Centrum Utrecht (UMC Utrecht). Jaardocument 2016, 2017.

[295] PT Vanberkel, RJ Boucherie, EW Hans, JL Hurink, and N Litvak. A survey of health care models that encompass multiple departments. Technical report, Department of Applied Mathematics, University of Twente, 2009.

[296] PT Vanberkel, RJ Boucherie, EW Hans, JL Hurink, WAM Van Lent, and WH Van Harten. An exact approach for relating recovering surgical patient work-

load to the master surgical schedule. *Journal of the Operational Research Society*, 62(10):1851–1860, 2011.

[297] PT Vanberkel, RJ Boucherie, EW Hans, JL Hurink, and N Litvak. Efficiency evaluation for pooling resources in health care. *OR Spectrum*, 34(2):371–390, 2012.

[298] M Vanhoucke and B Maenhout. Nspliba nurse scheduling problem library: A tool to evaluate (meta-) heuristic procedures. In *Operational research for health policy: making better decisions, proceedings of the 31st annual meeting of the working group on operations research applied to health services*, pages 151–165, 2007.

[299] M Vanhoucke and B Maenhout. On the characterization and generation of nurse scheduling problem instances. *European Journal of Operational Research*, 196(2): 457–467, 2009.

[300] LE van der Vechte. Patient preferences for duration and planning of diagnosis and start of treatment in cancer: A quantitative study. Master's thesis, University of Twente, 2016.

[301] R Veenhuizen and A Tibben. Coordinated multidisciplinary care for huntington's disease. an outpatient department. *Brain Research Bulletin*, 80(4):192–195, 2009.

[302] R Veenhuizen, B Kootstra, W Vink, J Posthumus, P van Bekkum, M Zijlstra, and J Dokter. Coordinated multidisciplinary care for ambulatory huntington's disease patients. evaluation of 18 months of implementation. *Orphanet Journal of Rare Diseases*, 6(1):1, 2011.

[303] MF van der Velde, N Kortbeek, and N Litvak. Organizing multidisciplinary care for children with neuromuscular diseases. *Health Systems*, pages 1–17, 2017.

[304] A Venkitasubramanian, SD Roberts, and JA Joines. Object oriented framework for healthcare simulation. In *Winter Simulation Conference (WSC), 2015*, pages 1436–1446. IEEE, 2015.

[305] I Vermeulen, S Bohte, K Somefun, and H La Poutré. Multi-agent pareto appointment exchanging in hospital patient scheduling. *Service Oriented Computing and Applications*, 1(3):185–196, 2007.

[306] IB Vermeulen, SM Bohte, SG Elkhuizen, PJM Bakker, and H La Poutré. Decentralized online scheduling of combination-appointments in hospitals. In *ICAPS*, pages 372–379, 2008.

[307] IB Vermeulen, SM Bohte, SG Elkhuizen, H Lameris, PJM Bakker, and H La Poutré. Adaptive resource allocation for efficient patient scheduling. *Artificial Intelligence in Medicine*, 46(1):67–80, 2009.

[308] SE Vernon. Continuous throughput rapid tissue processing revolutionizes histopathology workflow. *Laboratory Medicine*, 36(5):300–302, 2005.

[309] S Villa, A Prenestini, and I Giusepi. A framework to analyze hospital-wide patient

flow logistics: Evidence from an italian comparative study. *Health Policy*, 115(2): 196–205, 2014.

[310] E Visser, AG Leeftink, PSN van Rossum, S Siesling, R van Hillegersberg, and JP Ruurda. Waiting time from diagnosis to treatment has no impact on survival in patients with esophageal cancer. *Annals of Surgical Oncology*, 23:2679, 2016.

[311] E Visser, PSN van Rossum, AG Leeftink, S Siesling, R van Hillegersberg, and JP Ruurda. Impact of diagnosis-to-treatment waiting time on survival in esophageal cancer patients: A population-based study in The Netherlands. *European Journal of Surgical Oncology*, 43(2):461–470, 2017.

[312] MRM Visser, JJB van Lanschot, J van der Velden, JJ Kloek, DJ Gouma, and MAG Sprangers. Quality of life in newly diagnosed cancer patients waiting for surgery is seriously impaired. *Journal of Surgical Oncology*, 93(7):571–577, 2006.

[313] S Vlah, Z Lukač, and J Pacheco. Use of vns heuristics for scheduling of patients in hospital. *Journal of the Operational Research Society*, 62(7):1227–1238, 2011.

[314] MBV Rouppe van der Voort. *Optimising delays in access to specialist outpatient clinics*. Maastricht University, 2014.

[315] M van de Vrugt, RJ Boucherie, TJ Smilde, M de Jong, and M Bessems. Rapid diagnoses at the breast center of Jeroen Bosch Hospital: a case study invoking queueing theory and discrete event simulation. *Health Systems*, pages 1–13, 2016.

[316] K Walshe. Pseudoinnovation: the development and spread of healthcare quality improvement methodologies. *International Journal for Quality in Health Care*, 21(3):153–159, 2009.

[317] WY Wang and D Gupta. Adaptive appointment systems with patient preferences. *Manufacturing & Service Operations Management*, 13(3):373–389, 2011.

[318] T Wauters, J Kinable, P Smet, W Vancroonenburg, G Vanden Berghe, and J Verstichel. The multi-mode resource-constrained multi-project scheduling problem. *Journal of Scheduling*, 19(3):271–283, 2016.

[319] JD Welch and NTJ Bailey. Appointment systems in hospital outpatient departments. *The Lancet*, 259(6718):1105–1108, 1952.

[320] G Werker, A Sauré, J French, and S Shechter. The use of discrete-event simulation modelling to improve radiation therapy planning processes. *Radiotherapy and Oncology*, 92(1):76–82, 2009.

[321] J Whittle, G Schectman, N Lu, B Baar, and MF Mayo-Smith. Relationship of

scheduling interval to missed and cancelled clinic appointments. *The Journal of Ambulatory Care Management*, 31(4):290–302, 2008.

[322] L Wiesche, M Schacht, and B Werners. Strategies for interday appointment scheduling in primary care. *Health Care Management Science*, pages 1–16, 2016.

[323] JP Womack and DT Jones. *Lean thinking: banish waste and create wealth in your corporation*. Simon and Schuster, 2010.

[324] C Zacharias and M Pinedo. Appointment scheduling with no-shows and overbooking. *Production and Operations Management*, 23(5):788–801, 2014.

[325] L Zhao, CF Chien, and M Gen. A bi-objective genetic algorithm for intelligent rehabilitation scheduling considering therapy precedence constraints. *Journal of Intelligent Manufacturing*, pages 1–16, 2015.

[326] ME Zonderland and J Timmer. Optimal allocation of mri scan capacity among competing hospital departments. *European Journal of Operational Research*, 219 (3):630–637, 2012.

[327] ME Zonderland, F Boer, RJ Boucherie, A de Roode, and JW van Kleef. Redesign of a university hospital preanesthesia evaluation clinic using a queuing theory approach. *Anesthesia & Analgesia*, 109(5):1612–1621, 2009.

[328] ME Zonderland, RJ Boucherie, and A Al Hanbali. Appointments in care pathways: The geo^x/D/1 queue with slot reservations. *Queueing Systems*, 79(1): 37–51, 2015.

Acronyms

BCI	Batch Completion Interval
BCM	Batch Completion Moment
CHI	General surgery
CHOIR	Center for Healthcare Operations Improvement and Research
DC	Difference of Convex
DES	Discrete Event Simulation
EDD	Earliest Due Date
ENT	Ear-Nose-Throat
EYE	Ophthalmology
FCFS	First Come First Serve
GA	Genetic Algorithm
GI	Gastro Intestinal physician
GYN	Gynecology
GP	General Practitioner
HFPB	Hybrid Flow-shop with Parallel Batching
HFS	Hybrid Flow-Shop
HPB	Hepato-Pancreato-Biliary
ICU	Intensive Care Unit
ILP	Integer Linear Programming
ITAT	Intradepartmental Turnaround Time
KPI	Key Performance Indicator
LMS	Laboratory Management System
LOS	Length of Stay
LPT	Longest Processing Time
MDP	Markov Decision Process
MILP	Mixed Integer Linear Program
MIP	Mixed Integer Program
MPSM	Managerial Problem-Solving Method
MSE	Mean Squared Error
MSS	Master Surgery Schedule
MTM	Multi-disciplinary Team Meeting
NEU	Neurology and neurosurgery
NP	Nurse Practitioner
OLO	Unexpectedly Long Length of Stay
OM	Operations Management
ONC	Oncology

OR	Operations Research
ORT	Orthopedics
PDCA	Plan-Do-Check-Act
PLA	Plastic surgery
PSPLIB	Project Scheduling Problem Library
RCPSP	Resource Constrained Project Scheduling Problem
RIVM	Rijksinstituut voor Volksgezondheid en Milieu
RT	Radiotherapist
RTAT	Registration Turnaround Time
SAA	Sample Average Approximation
SIP	Stochastic Integer Programming
SPT	Shortest Processing Time
TAT	Turnaround Time
THO	Thoracic surgery
TOC	Theory Of Constraints
TQM	Total Quality Management
UCC	Utrecht Cancer Center
UMC Utrecht	University Medical Center Utrecht
URO	Urology
VBH	Value Based Healthcare
VRP	Vehicle Routing Problem
WKZ	Wilhelmina Kinder Ziekenhuis – UMC Utrecht's child hospital

Summary

In the Netherlands, patients receive a high quality of care. Research shows that successful care pathways exist for several types of cancer. However, patients might not receive this care to their full advantage if they have to wait.

The objective of this thesis is to improve the processes involved in the care pathway of cancer patients. An integrated approach towards the optimization of oncology processes is essential for the delivering of high quality cancer care. Herein, the whole cancer care chain is jointly optimized, as patients go through a patient journey that does not stop after diagnosis or treatment. This is reflected in the build-up of this thesis, which follows the steps of a cancer patient's care journey. Not only the entire cancer care chain, but also non-cancer care is important to take into account when optimizing oncology processes. As oncology care requires many shared resources, negative consequences of reserving capacity for the remaining patient population should be prevented.

To maximize the probability of improving patient care, healthcare Operations Management (OM) research not only includes theoretical, but also practical considerations in the research design. Furthermore, an integrated view not solely focuses on improving the quality of care, but simultaneously optimizes the quality of work and the productivity as well. This ensures that besides the best possible care for patients, a healthy work environment for employees is offered, from which as many patients as possible can benefit.

Part I Introduction

Part I presents the background information to this thesis.

Chapter 1 explains the stages in the care process of cancer patients, and provides more details on the various Operations Research (OR) methodologies that are applied in this thesis.

Chapter 2 provides an overview of the literature in multi-disciplinary appointment planning in healthcare. Multi-disciplinary planning is an emerging research field, which has many applications in healthcare, with similar underlying planning characteristics. We identify multiple fields to classify the literature upon. These fields relate to the system characteristics, decision characteristics, and applicability. The relevant papers for each of these fields are discussed, which provides a broad and thorough overview of the present research, and guides readers towards identifying the applicable literature for their research based on the characteristics of their problem. Furthermore, we disclose research gaps and present open challenges for further research.

Part II Diagnostics

One third of the Dutch population gets diagnosed with cancer during their lives. Many of these patients are offered a rapid diagnostics trajectory, which enables them to get a diagnosis within a short period of time. This requires flexibility from the cooperating disciplines, such as the department of pathology. Part II focuses on the operations in the histopathology laboratory, where the evaluation of possible cancerous tissue interrupts the regular workflow, causing an increase in throughput time and workload.

In *Chapter 3* we develop a 2-phased decomposition approach to improve the histopathology operations, which involves manual processes on single items, as well as automated batched processes. Our approach ensures that the throughput time of tissue samples in the laboratory are minimized, which enables fast diagnoses for patients, and yields a leveled workload for the technicians. In the first phase, a MILP equally divides the completion times of the batches in order to reduce the peaks of physical work available in the laboratory. The second phase minimizes the tardiness of orders using a list scheduling algorithm. We show that the resulting schedules can lead to a 50% reduction in workload, and a 20% reduction in turnaround times.

Chapter 4 continues with the model of Chapter 3, to analyze the effect of several interventions in the laboratory on the turnaround time of the tissue samples. It appears that especially the starting times of the tissue processors (multiple large batching processors in the laboratory that automate part of the processes in the laboratory) should be shifted towards specific moments during the day. These shifted starting times, combined with earlier starting shifts, can result in up to 25% decrease in turnaround time in the histopathology laboratory. UMC Utrecht has implemented the recommendations following from the Part II chapters in their histopathology workflow.

Part III Outpatient clinic

Cancer patients have multiple outpatient clinic consultations, not only for diagnostics, but also to discuss the treatment plan, and for follow-up consultations. Therefore, the planning of the outpatient clinic largely influences the patients' waiting and access time throughout their care pathway. Part III focuses on the planning decisions in the outpatient clinic.

Chapter 5 analyzes the optimal booking horizon for an outpatient clinic, given their patient characteristics. This shows a trade-off has to be made between the booking horizon and the probability that patients cancel or do not show up for their appointment. Depending on the specific parameter settings of a clinic, it might be beneficial for follow-up patients, who need a yearly checkup appointment, to postpone the moment of scheduling next years appointment, instead of immediately scheduling the next appointment for next year.

Chapter 6 analyzes the organization of a specialized oncology clinic, where patients enter with confirmed diagnoses that are in need of a treatment plan based on a multi-disciplinary approach during a one day visit. As the treatment plan is not yet known when the patients arrive at the clinic, we develop

a stochastic programming model to design a blueprint schedule for all involved specialists. Using this approach, robust blueprint schedules are found for one of UMC Utrecht's specialized oncology clinics.

Chapter 7 continues the work of Chapter 6 for the multi-disciplinary oncology clinics at the operational level of control. Using an extensive computer simulation model, various planning and scheduling rules are evaluated to support planners how to control the daily operations of these clinics. It is shown that a trade-off between waiting time and overtime should be made. Especially the invitation strategy and the routing rules have an impact on these performance indicators.

Part IV Treatment

Part IV focuses on the treatment of cancer patients. Surgery is one of the main treatment modalities for patients with a curable cancer. The planning of these surgeries is challenging, as the duration is hard to predict.

Chapter 8 introduces a case mix classification, which describes the volume and properties of the surgery types of a specialty. In order to compare the operating theater performance of the oncology specialty with other specialties, the specialties' underlying case mixes can be classified using this plot. Furthermore, a benchmark set of 20,880 instances is developed using a novel instance generation procedure, to compare the performance of surgery scheduling algorithms given the diverse range of case mixes.

Part V Conclusions

Part V focuses on the lessons learned from the presented research, and identifies trends and directions for further research.

In *Chapter 9* we analyze how to make an impact in practice with research in the field of OM/OR in healthcare. We explain the CHOIR ecosystem, which is based on the researcher-in-residence model, and discuss our experiences. Furthermore, we identify under which conditions researchers should perform their research, to maximize probability of having true impact:

First, a project can only be successful when scientific and healthcare people form a project team together. This means that scientific staff is introduced to the healthcare environment, and that healthcare staff is introduced to the engineering approach. Second, ensure that projects do not rely on a single person in the healthcare organization, but get the whole department or people from within the whole organization involved. Third, combine top-down and bottom-up approaches. This way, there is pressure to change the current situation, and there is commitment for the project from both front-line staff and administration. Fourth, a project should be a balance between theory and practice. The systematic problem solving approach adds value by not solving consequences, but the root cause of the problems. Furthermore, OM/OR approaches help organizations to improve their organization of processes using theoretically sound solutions. Note that when decisions in hospitals involve many people, solutions should be easy to implement. In automated systems, more involved solutions are appropriate. Fifth, clearly distinguish that researchers support in decision

making, and end users/management are the decision makers. This awareness ensures that end users are involved, and increases the probability that real impact is made.

Chapter 10 provides an outlook to this thesis. The current developments in oncology ask for specialization, personalization, and centralization of care. This requires an advanced organization of processes, which are not only multi-disciplinary, but also multi-organizational.

Samenvatting

In Nederland ontvangen patiënten een hoge kwaliteit van zorg. Onderzoek toont aan dat er voor verschillende typen kanker zeer goede zorgtrajecten ontwikkeld zijn. Maar patiënten kunnen deze zorg niet ten volste ontvangen wanneer ze moeten wachten.

In dit proefschrift onderzoeken we hoe we de processen in de kankerzorg kunnen verbeteren. Het is essentieel om dit geïntegreerd te benaderen om hoge kwaliteit van kankerzorg te bieden. Hierbij wordt het gehele zorgtraject gezamenlijk geoptimaliseerd, aangezien patiënten door een zorgtraject gaan dat niet stopt na de diagnose of behandeling. Deze geïntegreerde aanpak is terug te zien in de opbouw van dit proefschrift, waarin de stappen van een kankerpatiënt door de kankerzorg gevolgd worden. Het is niet alleen belangrijk om de kanker zorgketen in ogenschouw te nemen in de optimalisatie van processen in de oncologie, ook moet de zorg voor niet-oncologische patiënten meegenomen worden. Omdat de oncologische zorg veel gebruik maakt van gedeelde resources, moeten negatieve gevolgen van gereserveerde capaciteit voor de overige patiënten voorkomen worden.

Om de kans op een verbeterde patiëntzorg te maximalizeren, worden niet alleen theoretische overwegingen in onderzoek naar Operations Management (OM) in de zorg meegenomen, maar ook praktische overwegingen. Daarnaast vraagt een geïntegreerde aanpak niet alleen om een focus op een verbetering van de kwaliteit van zorg, maar om een focus waarbij tegelijkertijd ook de kwaliteit van werk en de productiviteit verbeterd worden. Dit zorgt ervoor dat de best mogelijke zorg voor patiënten wordt geboden, samen met een gezonde werkomgeving voor medewerkers, waarvan zoveel mogelijk patiënten gebruik kunnen maken.

Deel I Introductie
Deel I beschrijft de achtergrond van dit proefschrift.

Chapter 1 licht de verschillende stappen in het zorgproces van een kankerpatiënt toe. Daarnaast geeft het meer informatie over de verschillende Operations Research (OR) technieken die zijn toegepast in dit proefschrift.

Chapter 2 geeft een overzicht van de huidige literatuur met betrekking tot multi-disciplinaire afsprakenplanning in de gezondheidszorg. Multi-disciplinaire afspraken planning is een opkomend onderzoeksgebied, met veel toepassingen in de gezondheidszorg. Voor deze toepassingen geldt dat de onderliggende planningskarakteristieken vaak vergelijkbaar zijn. We presenteren een indeling van de literatuur, gerelateerd aan de systeemkarakteristieken, belissingskarakteristieken, en de toepassing. Door de relevante onderzoeken in dit veld te beschrijven, geven

we een breed en diepgaand overzicht van de huidige literatuur, en bieden we de lezers een handvat om de voor hun onderzoek relevante literatuur te identificeren gebaseerd op de karakteristieken van hun probleem. Daarnaast beschrijven we de hiaten in de literatuur, en presenteren we mogelijkheden voor toekomstig onderzoek.

Deel II Diagnostiek
Een derde van de Nederlandse populatie wordt gediagnosticeerd voor kanker gedurende hun leven. Veel van deze patiënten wordt een sneldiagnostiektraject aangeboden, waarmee ze hun diagnose zo snel mogelijk krijgen. Hiervoor is flexibiliteit van de betrokken disciplines nodig, zoals van de afdeling pathologie. Deel II focust op de activiteiten in het histopatologie laboratorium, waar spoedopdrachten voor kankerpatiënten de reguliere workflow interrumperen. Dit creeert een ongewenste verhoging van de doorlooptijd van weefsels en van de werkdruk voor medewerkers.

In *Hoofdstuk 3* ontwikkelen we een 2-fasen methode om de logistiek in het histopatologie laboratorium te verbeteren, waarbij rekening wordt gehouden met handmatige processen voor losse weefsels en geautomatiseerde processen voor batches van weefsels. Onze methode minimaliseert de doorlooptijd van alle weefsels in het laboratorium, waardoor snelle diagnostiek voor patiënten kan worden geleverd. Ook creert het een gelijkmatig verdeelde werkdruk voor de technici. In de eerste fase, verdeelt MILP de tijd dat de batches klaar zijn over de dag, zodanig dat de stapels werk in het laboratorium gereduceerd worden. In de tweede fase minimaliseert een list scheduling algoritme de kans dat weefsels te laat klaar zijn. De uitkomsten van dit onderzoek laten zien dat met de nieuwe planning, een 50% reductie in werkdruk en een 20% reductie in doorlooptijd behaald kan worden. *Hoofdstuk 4* vervolgt met het model van Hoofdstuk 3, om het effect van verscheidene interventies op de doorlooptijd van de weefsels in het laboratorium te testen. Het grootste effect kan behaald worden door de doorvoermachines (meerdere grote batchmachines die een gedeelte van de processen in het laboratorium automatiseren) op specifieke momenten in de dag te starten. Deze starttijden, samen met een vervroegde start van een deel van het personeel, kan resulteren in een 25% reductie in doorlooptijd in het histopatologie laboratorium. UMC Utrecht heeft de aanbevelingen uit Deel II van dit proefschrift geimplementeerd in hun dagelijkse werkprocessen.

Deel III Polikliniek
Kankerpatiënten hebben verschillende poliklinische consulten, niet alleen voor diagnostiek, maar ook om het behandelplan te bespreken of voor follow-up. Daarom beinvloedt de planning van de polikliniek op veel plekken de wacht- en toeganstijd voor de patiënt. Deel III focust op planningsbeslissingen in de polikliniek.

Hoofdstuk 5 analyseert de optimale boekingshorizon voor een polikliniek, gegeven de patiëntkarakteristieken van deze kliniek. De resultaten laten zien dat een afweging moet worden gemaakt tussen de boekingshorizon, en de kans dat

een patiënt zijn of haar afspraak afzegt of hier niet voor op komt dagen. Afhankelijk van de specifieke parameters van een kliniek, kan het voordelig zijn om voor follow-up patiënten, die bijvoorbeeld een jaarlijkse onder controle staan, het moment van het plannen van de afspraak voor volgend jaar uit te stellen, in plaats van direct een volgende afspraak te maken voor volgend jaar.

Hoofdstuk 6 analyseert de organisatie van een gespecialiseerde oncologische polikliniek, waarin patiënten met een kankerdiagnose via een multi-disciplinaire benadering hun behandelplan krijgen gedurende één ziekenhuisbezoek. Aangezien het behandelplan nog niet bekend is wanneer de patiënten in de kliniek arriveren, ontwikkelen we een stochastisch model dat de blauwdruk van de agenda voor de betrokken specialisten bepaalt. Met deze methode zijn robuste afsprakenschema's gevonden voor een van UMC Utrechts gespecialiseerde kankerklinieken.

Hoofdstuk 7 borduurt voort op het werk van Hoofdstuk 6 voor de multidiscipinaire polikliniek van op een operationeel niveau. Door middel van een uitgebreid computersimulatiemodel zijn verschillende planningsregels geevalueerd om planners te ondersteunen in de dagelijkse besturing van de polikliniek. We laten zien dat een afweging gemaakt moet worden tussen wachttijd en overtijd. Met name het uitnodigingsbeleid en de routering door de kliniek hebben een grote invloed op deze prestatie-indicatoren.

Deel IV Behandeling

Deel IV focust op de behandeling van kankerpatiënten. Het operatief verwijderen van de tumor is de meest gekozen curatieve behandelingswijze. De planning van deze operaties is een uitdaging, aangezien de duur van de operaties moeilijk te voorspellen is.

Hoofdstuk 8 introduceert een case mix classificatie, die het volume en de eigenschappen van operaties van een specialisme beschrijft. Om een vergelijking te maken van de prestatie van specialismen, zoals de oncologie, kan de onderliggende case mix van een specialisme worden geclassificeerd in een plot. Daarnaast is een benchmarkset van 20.880 instanties ontwikkeld, met gebruik van een slimme procedure die instanties genereert. Hiermee kan de prestatie van algoritmes voor operatieplanning vergeleken worden, gegeven de diversiteit aan case mixen.

Deel V Conclusies

Deel V presenteert de opgedane kennis van het onderzoek in dit proefschrift, en identificeert trends en mogelijkheden voor vervolgonderzoek.

In *Hoofdstuk 9* analyseren we hoe een impact in de praktijk kan worden gemaakt met het onderzoek naar OM/OR in de zorg. We introduceren het CHOIR ecosysteem, gebaseerd op het 'researcher-in-residencemodel', en delen onze ervaringen met praktijkgericht onderzoek. Daarnaast identificeren we onder welke omstandigheden onderzoeker hun onderzoek moeten doen, om de kans op impact in de praktijk te verhogen: Allereerst kan een project alleen succesvol zijn als het uitgevoerd wordt door een projectteam met onderzoekers en zorgprofessionals. Dit betekent dat de onderzoekers geïntroduceerd worden in de zorgomgeving, en dat de engineering-aanpak geïntroduceerd wordt aan zorgprofessionals. Ten

tweede staat of valt een project door de betrokkenheid van de hele afdeling, of meerdere mensen vanuit de hele organisatie. Één persoon is dus niet genoeg. Ten derde, pas een gecombineerde toepassing van top-down en bottom-up benaderingen toe. Hierdoor is er druk om de huidige situatie te veranderen, en inzet voor het project van zowel de werkvloer als het management. Ten vierde, vind een goede balans tussen theorie en praktijk. Het systematisch oplossen van logistieke problemen voegt waarde toe door niet te focussen op het oplossen van gevolgen, maar door het kernprobleem op te lossen. Daarnaast kan een OM/OR aanpak organisaties helpen om hun processen te verbeteren door theoretisch solide oplossingen. Let op dat als veel mensen betrokken zijn bij beslissingen in het ziekenhuis, de oplossingen simpel te implementeren moeten zijn. In geautomatiseerde systemen zijn meer ingewikkelde oplossingen op zijn plaats. Als laatste moeten onderzoekers beseffen dat zij ondersteunen in de besluitvorming, maar dat de besluiten gemaakt worden door de eindgebruiker en/of het management. Dit besef zorgt ervoor dat de eindgebruiker betrokken wordt bij het proces, en vergroot de kans op impact in de praktijk.

Hoofdstuk 10 geeft een vooruitblik naar aanleiding van dit proefschrift. De huidige ontwikkelingen in de oncologie vragen om specialisatie, personalisatie, en centralisatie van de zorg. Dit vraagt om een geavanceerde organisatie van processen, die niet alleen multi-disciplinair zijn, maar ook meerdere organisaties betrekt.

About the author

Gréanne Leeftink was born in Drachten, Smallingerland, the Netherlands, on August 30, 1991. In 2000 she started her higher education at Stedelijk Gymnasium Nijmegen, and she obtained her diploma for preparatory university education (in Dutch: Gymnasium) from Gomarus College Groningen in 2009.

Gréanne studied Industrial Engineering and Management at the University of Twente in Enschede. She visited the University of Toronto in 2012, where she performed a research project in the ambulatory laboratory of Mount Sinai Hospital. Gréanne obtained her bachelor's degree with honors from the University of Twente in 2012.

During her masters program, Gréanne specialized in Production and Logistics Management. She performed her master's thesis project at the histopathology laboratory of University Medical Center Utrecht (UMC Utrecht), which was the basis of Chapter 3 and Chapter 4 of this thesis. In 2014 Gréanne defended her master thesis cum laude with honors. For this thesis, she received the second prize in the Material Handling Master thesis competition.

During her studies, Gréanne wrote a research proposal 'Rapid diagnostics for cancer? Yes, we can!', together with Prof.dr.ir. Erwin W. Hans and dr.ir. Ingrid M.H. Vliegen. With this proposal, she was awarded an individual talent grant of the Netherlands Organization for Scientific Research (in Dutch: NWO). With this grant, she started her Ph.D. in September 2014 at the Center for Healthcare Operations Improvement and Research (CHOIR) of the University of Twente, within the department of Industrial Engineering and Business Information Systems. For her Ph.D. research, she got also positioned at UMC Utrecht Cancer Center.

Gréanne was awarded a travel grant from the Prins Bernhard Cultuurfonds - Data Piet Fonds to visit the Health Care Systems Engineering lab of Mayo Clinic in Rochester, MN, USA from September until December 2016. The joint work with Dr. Kalyan S. Pasupathy and Dr. Mustafa Y. Sir formed the basis of Chapter 5 of this thesis.

Following her Ph.D., Gréanne will be appointed as an assistant professor at the University of Twente and the CHOIR research center. She will continue her research in healthcare Operations Management, with a focus on the planning of integrated processes.

List of publications

E. Visser, A.G. Leeftink, P.S.N. van Rossum, S. Siesling, R. van Hillegersberg, and J.P. Ruurda. Waiting time from diagnosis to treatment has no impact on survival in patients with esophageal cancer. *Annals of Surgical Oncology*, 23(8):2679-2689, 2016.
(Basis for Chapter 1.)

E. Visser, P.S.N. van Rossum, A.G. Leeftink, S. Siesling, R. van Hillegersberg, and J.P. Ruurda. Impact of diagnosis-to-treatment waiting time on survival in esophageal cancer patients A population-based study in The Netherlands. *European Journal of Surgical Oncology*, 43(2):461-470, 2017.
(Basis for Chapter 1.)

A.G. Leeftink, I.A. Bikker, I.M.H. Vliegen, and R.J. Boucherie. Multi-disciplinary appointment planning - a review. *Submitted.*
(Basis for Chapter 2.)

A.G. Leeftink, R.J. Boucherie, E.W. Hans, M.A.M. Verdaasdonk, I.M.H. Vliegen, and P.J. van Diest. Batch scheduling in the histopathology laboratory. *Flexible Services and Manufacturing Journal*, `https://doi.org/10.1007/s10696-016-9257-3`, 2017.
(Basis for Chapter 3.)

A.G. Leeftink, R.J. Boucherie, E.W. Hans, M.A.M. Verdaasdonk, I.M.H. Vliegen, and P.J. van Diest. Histopathology laboratory operations analysis and improvement. In R. Grlitz, V. Bertsch, S. Caton, N. Feldmann, P. Jochem, M. Maleshkova, and M. Reuter-Oppermann (Eds.), *Proceedings of the First Karlsruhe Service Summit Research Workshop - Advances in Service Research*, pp. 51-63, 2015. Karlsruhe: Karlsruher Institut fur Technologie (KIT) Scientific Publishing.
(Basis for Chapter 3.)

A.G. Leeftink, R.J. Boucherie, E.W. Hans, M.A.M. Verdaasdonk, I.M.H. Vliegen, and P.J. van Diest. Predicting turnaround time reductions of the diagnostic track in the histopathology laboratory using mathematical modelling. *Journal of Clinical Pathology*, 69(9):793-800, 2016.
(Basis for Chapter 4.)

List of publications

A.G. Leeftink, M.G. Martinez, E. Sisikoglu Sir, E.W. Hans, M.Y. Sir, and K.S. Pasupathy. Scheduling interval optimization in healthcare clinics under patient no-show and cancellations behavior. *Submitted.*
(Basis for Chapter 5.)

A.G. Leeftink, I.M.H. Vliegen, and E.W. Hans. Stochastic integer programming for multi-disciplinary outpatient clinic planning. *Health Care Management Science*, `https://doi.org/10.1007/s10729-017-9422-6`, 2017.
(Basis for Chapter 6.)

A.G. Leeftink and E.W. Hans. Case mix classification and a benchmark set for surgery scheduling. *Journal of Scheduling*, `https://doi.org/10.1007/s10951-017-0539-8`, 2017.
(Basis for Chapter 8.)

Beta
Research School for Operations Management and Logistics

UNIVERSITY OF TWENTE.
Faculty of Behavioural, Management & Social Sciences
Department of Industrial Engineering & Business Information Systems

UMC Utrecht

CTIT

NWO

CHOIR

The access to cancer diagnostics and cancer treatment is not the same for all types of cancer patients. Furthermore, the resources involved in these processes are costly and scarce. Long access and waiting times to diagnostics and treatment can cause increased anxiety of patients.

The goal of this thesis is to improve the quality and efficiency of (multi-disciplinary) cancer care processes. We develop new planning and control approaches to optimize the organization of multiple shared resources involved, so that employees experience a leveled workload, and access to diagnostics and treatment is equally divided over and optimized for all patient types. To develop these approaches we use Operations Management/ Operations Research techniques. Furthermore, we apply the outcomes through case studies in UMC Utrecht (NL) and Mayo Clinic (USA), and analyze the critical success factors for making an impact in practice.

ISBN 978-90-365-4411-5