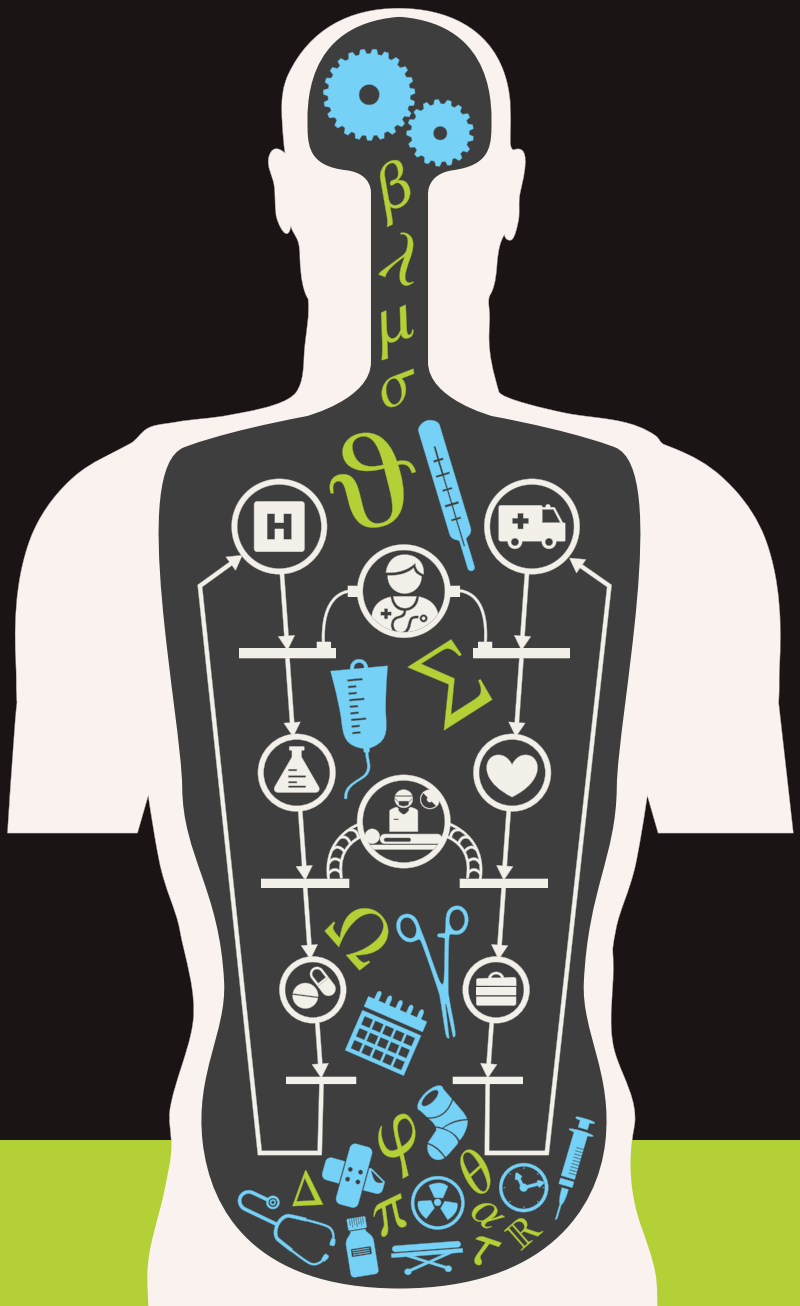
QUALITY-DRIVEN EFFICIENCY IN HEALTHCARE

NIKKY KORTBEEK



QUALITY-DRIVEN EFFICIENCY IN HEALTHCARE

Nikky Kortbeek

Dissertation committee

Chairman & secretary:	Prof. dr. ir. A.J. Mouthaan
Promotors:	Prof. dr. R.J. Boucherie Prof. dr. P.J.M. Bakker
Members:	Prof. dr. S.C. Brailsford Prof. dr. M.W. Carter Prof. dr. N.M. van Dijk Dr. ir. E.W. Hans Prof. dr. J.L. Hurink Prof. dr. I.N. van Schaik Prof. dr. P.G. Taylor

Ph.D. thesis, University of Twente, Enschede, the Netherlands
Center for Telematics and Information Technology (No. 12-231, ISSN 1381-3617)
Beta Research School for Operations Management and Logistics (No. D162)
Center for Healthcare Operations Improvement and Research

This research was financially supported by the Dutch Technology Foundation STW by means of the project 'Logistical Design for Optimal Care' (No. 08140)

Publisher: Gildeprint Drukkerijen, Enschede, the Netherlands
Cover design: Bundelmedia, Beverwijk, the Netherlands

ISBN 978-90-365-3428-4
DOI 10.3990/1.9789036534284

QUALITY-DRIVEN EFFICIENCY IN HEALTHCARE

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
Prof. dr. H. Brinksma,
volgens besluit van het College voor Promoties,
in het openbaar te verdedigen
op vrijdag 23 november 2012 om 14.45 uur

door

Nikky Kortbeek

geboren op 1 november 1983
te Beverwijk, Nederland

Dit proefschrift is goedgekeurd door de promotores:

Prof. dr. R.J. Boucherie, en
Prof. dr. P.J.M. Bakker

"Knowing is not enough; we must apply. Willing is not enough; we must do."

— Johann Wolfgang von Goethe

Voorwoord

Tijdens het afronden van mijn studie kwam het moeten beantwoorden van de welbekende existentiële vraag “wat wil je later worden?” snel dichterbij. De keuze om promotieonderzoek te gaan doen, is allesbehalve kiezen voor de weg van de minste weerstand. Toch heb ik er geen seconde spijt van gehad. Afgezien van het feit dat het gewoon leuk is, hoop ik met dit proefschrift een bijdrage te kunnen leveren aan het bekender maken van de maatschappelijke waarde van de wiskunde.

Dit proefschrift draagt mijn naam, maar is allesbehalve een individueel resultaat. Voortkomend uit mijn passie voor het samenbrengen van wetenschap en praktijk, hebben een heel aantal mensen met verschillende achtergronden bijgedragen aan de totstandkoming van dit proefschrift. Zonder de illusie te koesteren uitputtend te kunnen zijn, gebruik ik deze plaats om enkele personen expliciet te bedanken.

Nico van Dijk, als inspirator en leermeester heb jij grote invloed gehad op mijn keuze om promotieonderzoek te gaan doen. Jij liet mij mijn eerste stappen zetten in de academische wereld: eerst als student-assistent, daarna als docent en onderzoeker. Met jou schreef ik mijn eerste wetenschappelijke artikel, jij durfde het aan mij voor de klas te zetten om college te geven aan medestudenten, en om mij als onervaren onderzoeker als ‘expert’ naar de bloedbank te sturen. Door dit door jou getoonde vertrouwen groeide mijn geloof dat promoveren voor mij zou zijn weggelegd. Ook praktisch stond jij aan de basis door mij in contact te brengen met mijn (wat zou blijken) promotoren: Richard Boucherie en Piet Bakker.

Richard, vanaf de eerste dag voelde ik mij zeer thuis bij jouw gedrevenheid, scherpte, en directheid. Wij spreken elkaars taal. Ik besef dat ik het jou door mijn eigenwijsheid en daarmee gepaard gaande ongrijpbaarheid niet makkelijk heb gemaakt. Ik waardeer het dat je mij de ruimte liet mijn eigen keuzes liet maken, en erop vertrouwde dat het ergens toe zou leiden. Je bent voor mij een baken op de weg naar academische volwassenheid.

Piet, jouw vastberadenheid om de gezondheidszorg naar een hoger plan te tillen werkt inspirerend. Jouw kennis van de medische wereld en de bereidheid om als arts daarover geen blad voor de mond te nemen, hebben voor mij deze wereld geopend. Door jouw brede interesse sla je een brug tussen verschillende vakgebieden; daarmee ben je in mijn ogen een stuwende kracht voor een vooruitgaande samenleving. Ik beschouw onze gesprekken over hoe wiskundige resultaten uit te leggen aan zorgprofessionals als zeer waardevol. Mijn besluit om na mijn promotie aan de slag te gaan bij de door jou geleide afdeling KPI van het AMC geeft blijk van onze prettige samenwerking. Ons werk is nog niet af.

Ik wil de leden van de promotiecommissie bedanken, te beginnen met Erwin Hans. Erwin, jouw inhoudelijke bijdrage aan het tweede, derde en zesde hoofdstuk van dit proefschrift is onmiskenbaar. Daarnaast: heeft iemand jou wel eens verteld dat jouw enthousiasme aanstekelijk werkt? Ik heb veel opgestoken van jouw aanpak bij het samen begeleiden van studenten tijdens hun afstudeeronderzoek. Met name de vrijdagmiddagbesprekingen die werden afgesloten met het youtube-filmpje van de week staan in mijn geheugen. Ik ben ook Ton Mouthaam, Sally Brailsford, Michael Carter, Johann Hurink, Ivo van Shaik en Peter Taylor erkentelijk voor het nemen van zitting in mijn promotiecommissie.

I am thankful to Peter Taylor, Anneke Fitzgerald, Kate Hayes, and Terry Sloan for facilitating my research visit to Australia. Peter, thank you for inviting me to the University of Melbourne. I enjoyed our collaboration of which the results are partly reflected in Chapter 12 of this thesis. Anneke, Kate, and Terry, thank you for hosting me at the University of Western Sydney and Campbelltown Hospital. Anneke, the hospitality you showed by opening your house to us was heartwarming. My sincere apologies for, despite promises, not having proved to be able to sell your house during your holiday.

Ik wil ook mijn dank uitspreken aan de personen die een specifieke bijdrage hebben geleverd aan het onderzoek dat is beschreven in de verschillende hoofdstukken van dit proefschrift: Peter Hulshof (Hoofdstuk 2), Maartje Zonderland en Nelly Litvak (Hoofdstuk 3), Nelly Litvak, Marjan van der Velde, Ellen Dibbits, Bert Kiewiet en Liesbeth Flippo (Hoofdstuk 4), Aleida Braaksma, Kees Bijl, Henk Greuter, Frans Nollet en Gerhard Post (Hoofdstuk 5), Nelly Litvak, Niek Baer en Olaf Roukens (Hoofdstuk 6), Aleida Braaksma, Christian Burger, Ferry Smeenk, Chris Bakker en Reggie Smith (Hoofdstukken 7 en 8), en Erik van Ommeren (Hoofdstuk 12). Aan mijn collega's Andreas Fügener, Jelmer Kranenburg, Frank Mak, Jasper van Sambeek, Peter Vanberkel, Joost Veldwijk en Ingrid Vliegen wil ik zeggen: ons lopend onderzoek heeft dit proefschrift net niet gehaald, de invloed van mijn samenwerking met jullie is niettemin weerspiegeld in het huidige resultaat.

Ik bedank alle collega's van CHOIR, SOR en KPI. Hiervan wil ik er nog een aantal in het bijzonder wil noemen.

Nelly, je stelde de juiste vragen, en deed alles om te helpen bij het zoeken naar de juiste antwoorden. Zo ook toen je ons een stelling uit een Russisch wiskundeboek aandroeg, toen Maartje en ik toch echt dachten te zijn vastgelopen. Ik bewonder jouw vermogen om de beschrijving van een wiskundig model compact en kraakhelder op papier te zetten. Hier heb ik zeker mijn voordeel mee gedaan.

Peter (Hulshof), ik heb me wel eens afgevraagd of ons literatuuronderzoek er ooit was gekomen als we wisten waar we aan begonnen. Het is in ieder geval het hoofdstuk waar op de meeste verschillende plekken op deze wereld aan is gewerkt. Gedeeld perfectionisme maakte ons als team sterk, maar was ook onze zwakte. Ik heb genoten van alle discussies over kleine nuances in formuleringen. Mijn Engelse schrijfvaardigheid heeft er zeker van geprofiteerd. Ik hoop dat ons team nog eens in ere hersteld wordt, al was het maar om nog eens de Belgische horeca te trotseren.

Maartje (Zonderland), jij was degene die mij bij binnenkomst wegwijs maakte op de UT. Na jouw korte afwezigheid deed het me goed jou opnieuw te mogen verwelkomen als kamergenoot. Af en toe de deur dicht doen en even de grote boze buitenwereld met jou bespreken kan zo lekker opluchten. Het is eigenlijk jammer dat onze samenwerking beperkt is gebleven tot één project. Wat niet is, kan nog komen.

Aleida, ik had het genoegen jou te mogen begeleiden tijdens jouw afstudeeronderzoek. Ik was heel blij toen je daarna besloot collega-promovendus te worden bij de UT en het AMC. Zowel inhoudelijk als op persoonlijk vlak heeft het mijn promotietraject kleur gegeven. Het siert je dat je je bij momenten schijnbaar nog meer bekommerde om mijn deadline dan ikzelf. Ik zie er naar uit om onze samenwerking voort te zetten.

Peter (Vanberkel), you did the pioneering work being the first CHOIR PhD, from which all your successors, myself included, benefit. I also want to point out that you were the one who laid the theoretical foundation for Chapters 7 and 8.

Theresia, jouw masterclass figuren maken in Latex heeft zijn vruchten afgeworpen. Misschien spingt daarmee jouw invloed op dit proefschrift nog wel het meest in het oog.

Egbert, onze repeterende strijd om wie de hardste lach kan opwekken tijdens een presentatie op een wetenschappelijk congres is nog onbeslist. Ik daag je uit voor een volgende ronde.

Maartje (van de Vrugt), de week die wij samen doorbrachten in Beijing was enerverend. Ik blijf benieuwd of de muzikale taxichauffeur ons nog heeft opgenomen in zijn hall of fame.

Tot slot richt ik het woord tot mijn familie en vrienden. Dennis en Christiaan, geweldig dat jullie mij als paranimfen bijstaan bij de promotie. Vriendschap is niet vanzelfsprekend. Ik beloof weer wat vaker naar buiten te komen. Ab en Mariëtte, jullie zijn een voorbeeld voor velen, niet in de laatste plaats voor mij. Johanna, het vervult mij van trots een oma als jij te hebben. Timo en Lotte, jullie zijn een broer en zus om van te houden. Edith en Herman, jullie gaan voor mij door het vuur en dat maakt mij sterk.

Lieve Annika, wat ik later wil worden weet ik nog steeds niet, maar jij maakt dat ik weet wie ik nu wil zijn.

Nikky
Wijk aan Zee, oktober 2012

Contents

Part I

Introduction

CHAPTER 1

Research Motivation and Outline

During the upcoming decades, healthcare organizations face the challenge to deliver more patient care, of higher quality, and with less financial and human resources. The goal of this dissertation is, by developing operations research techniques, to help and guide healthcare professionals making their organizations future-proof.

1.1 Healthcare in the 21st century

During the 20th century, healthcare delivery has contributed to a striking worldwide health improvement. Despite its unmistakable benefits, the healthcare sector is under serious strain [466, 639]. Demand for and expenditures on healthcare increase steadily, as a result of ageing populations, technological developments, and increased medical knowledge. At the same time, patient expectations, competition between healthcare organizations, and labor shortages are rising. A joint effort is required by policy-makers, insurers, and care providers to fundamentally reconsider the way healthcare is delivered.

Since 1960, life expectancy has increased on average across countries of the OECD (Organisation of Economic Co-operation and Development) by more than 11 years, reaching nearly 80 years in 2009 [466]. Concurrently, the past 50 years have shown a steady rise in healthcare spending, which has tended to grow faster than Gross Domestic Products (GDP). In 1960, health spending among health systems in OECD countries accounted for under 4% of GDP on average. By 2009, this had risen to 9.6%, with many countries spending over 10% of GDP. Particularly in the United States, the health spending share of GDP grew rapidly from about 5% in 1960 to over 17% in 2009. The next highest country, allocating 12%, was the Netherlands.

The Netherlands is a striking example of a country facing tremendous healthcare challenges. The Dutch government is convinced of the urgency of the problem [445]. Without drastic measures being taken, the Netherlands Bureau for Economic Policy Analysis (CPB) predicts that the health spending share of GDP potentially grows to more than 30% in 2040 (see Figure 1.1) [109]. With more people demanding care and a workforce that is not expected to grow in size, the share of the working population employed in the healthcare sector is expected to increase sharply (see Figure 1.2). These developments will put under pressure other areas that drive society, like education, social security, and environmental welfare.

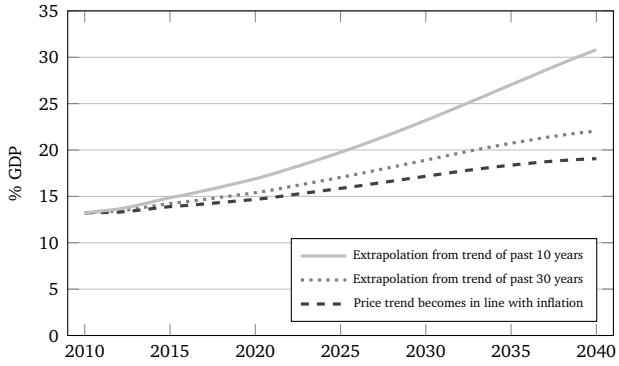


Figure 1.1: Predictions for total expenditure on health as share of GDP in the Netherlands (Source: The Netherlands Bureau for Economic Policy Analysis (CPB) [109]).

In an effort to break these trends, in 2006, the Dutch government changed the national healthcare system by introducing a limited form of competition [627]. As the system is still a 'work in progress', it is too early to tell whether the reforms can be considered as a success [515]. Whether the policy changes lead to improved quality, decreased costs, and increased innovation can only be fairly judged in the long term.

Performance levels of healthcare systems vary markable among high-income countries [77]. According to the OECD [466], the relationship between higher health spending per capita and higher life expectancy tends to be less pronounced as countries spend more on health. They conclude that the weak correlation at high levels of health expenditure suggests that there is room to improve the efficiency of health systems to ensure that the additional money spent on health brings about measurable benefits in terms of health outcomes. It is an observation that is shared by the World Health Organization (WHO), who state that opportunities to achieve more with the same resources exist in all countries [639]. They claim that, conservatively speaking, about 20–40% of resources spent on health are wasted through inefficiency.

Thus, with current efficiency levels being insufficient to keep healthcare affordable and accessible, let alone to be able to increase its quality, governments and healthcare providers must develop systems that deliver the best healthcare for the limited resources that are available. Where governments have to focus on effective policy-making and designing financial systems that provide the correct financial incentives, healthcare providers are responsible for decisions about clinical practice and the management of healthcare operations. This dissertation is directed to the level of the healthcare providers. Building from operations research techniques, and focusing on the management of operations, the aim of the research presented in this thesis is to contribute to a better understanding and functioning of healthcare delivery, and to support decision makers in realizing the best possible use of available resources.

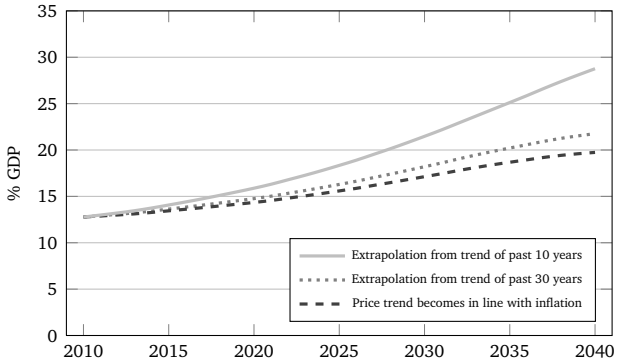


Figure 1.2: Predictions for share of Dutch workforce employed in health occupations (Source: The Netherlands Bureau for Economic Policy Analysis (CPB) [109]).

1.2 Quality-driven efficiency

Within a healthcare organization, professionals of different disciplines jointly organize healthcare delivery with the objective to provide high quality care using the limited resources that are available. The Institute of Medicine (IOM) outlines six specific aims that healthcare delivery must fulfil [325]. It must be *safe* (avoiding injuries to patients from the care that is intended to help), *effective* (providing services based on scientific knowledge to all who could benefit, and refraining from providing services to those not likely to benefit), *patient-centered* (providing care that is respectful of and responsive to individual patient preferences, needs, and values, and ensuring that patient values guide all clinical decisions), *timely* (reducing waits and sometimes harmful delays for both those who receive and those who give care), *efficient* (avoiding waste, including waste of equipment, supplies, ideas, and energy), and *equitable* (providing care that does not vary in quality because of personal characteristics such as gender, ethnicity, geographic location, and socioeconomic status).

Designing and organizing processes is referred to by the term 'planning and control'; it involves setting goals and deciding in advance what to do, how to do it, when to do it and who should do it. With the aim to achieve the goals formulated by the IOM, healthcare planning and control comprises multiple managerial functions, making medical, financial and resource decisions. This dissertation addresses the managerial function of resource capacity planning and control as defined in [273]: 'Resource capacity planning and control concerns the dimensioning, planning, scheduling, monitoring, and control of renewable resources (i.e., facilities, equipment and staff).' The research described contributes to the achievement of the necessary efficiency gains, while never losing sight of, in fact, while integrally improving on the various IOM quality dimensions. Thus, to achieve what is reflected by the title of this dissertation: quality-driven efficiency.

Planning and control has a rich tradition in manufacturing [611]. The nature of healthcare operations inhibits direct copying of successful industry practices, as it has certain distinctive characteristics [468, 483, 636]. Without being exhaustive, let us mention two prominent examples. First, because patients are part of the production process, the heterogeneity of patient's conditions and personalities makes the effectiveness of diagnosis and treatment outcomes strongly dependent on the individual patients. Therefore, standardization of operations is only possible to a limited extent [76]. Second, as a service industry, healthcare is produced and consumed simultaneously: care supply cannot be stored. Since the influence on care demand is limited and desired service levels are typically high, buffer capacity is required to cope with uncertain demand [271]. A certain degree of unused capacity must therefore be accepted, to keep accessibility on a sufficiently high level. Both these examples touch upon the issue of variability.

Variability is a concept inherently attached to healthcare operations. It complicates capacity planning and control. The challenge is to reduce variability when possible and deal with it when necessary. In that light, we can make a distinction between natural and artificial variability [399, 438]. Natural variability is the source of uncertainty that one has to deal with, as it is unavoidable (e.g., when it involves the number of patient presentations at an emergency department [252]), or even desirable (e.g., when it involves treatment customization [436]). Although sometimes ignored, it is often possible to proactively anticipate natural variability as it is generally to a certain degree predictable (e.g., seasonal demand patterns [483]). Artificial variability concerns variation that is undesirably created by deficiencies in planning and control (e.g., when elective clinical admissions cause unnecessary fluctuations in bed occupancy [99]), and thus should be prevented as much as possible. All studies in this dissertation contain elements addressing the challenge of reducing artificial variability and anticipating natural variability.

Realizing high-quality care delivery demands coordinated long-term, medium-term and short-term decision making. The planning and control decisions that have to be made are as diverse as numerous. In Chapter 2, we present a taxonomy along which we identify planning decisions in different areas of healthcare services and classify these in hierarchical levels. The taxonomy adopts the the four hierarchical (temporal) levels presented in the framework of [273], which applies the well-known breakdown of *strategic*, *tactical* and *operational* planning [17]. The operational level is subdivided in *offline* and *online* decision making, where *offline* reflects the in advance decision making and *online* the real time reactive decision making in response to events that cannot be planned in advance. The structured literature review that is performed in Chapter 2 based on the proposed taxonomy, exposes the importance of hierarchical alignment between strategic, tactical, and operational decision making. For example, meaningful surgical case scheduling (operational) can only be achieved when surgeon staffing levels are appropriate (tactical) and enough operating rooms are constructed (strategic). The research presented in the chapters that follow will reinforce the observation that recognizing and incorporating the hierarchical relations in decision making improves healthcare delivery performance.

In general, the clinical course is a highly fragmented process, because multiple clinicians of different departments or even organizations are involved in a patient's treatment. When there is a lack of coordination and collaboration between the actors within a care chain, the risk exists of clinical and logistical misalignment between consecutive treatment steps. This has negative consequences on patient outcomes, patient satisfaction, and resource utilization [102, 591]. Manifestations of logistical misalignment are for example excessive delays between treatment stages by which patients' conditions deteriorate while they spend time on a waiting list [486], and resources that are misused because patients cannot timely continue to consecutive treatment steps. The latter can occur within organizations, for instance when a patient in an intensive care bed waits for a bed at a general ward [635], but also between organizations, for example when a patient in a hospital bed waits for admission at a rehabilitation facility [410]. In challenging clinical misalignment, thereby avoiding under- and overtreatment, organizing care in closely cooperating multidisciplinary teams, covering the full range of physical, psychological, social, preventive, and therapeutic modalities, is emerging as a promising approach [407]. To conclude, in addition to alignment between hierarchical decisions, coordination and collaboration within a care chain is essential. The value of establishing clinical and logistical synergy is underlined by many of the chapters in this dissertation.

The final recurring theme in this dissertation is that of flexibility. Flexibility in resource capacity planning and control involves the ability to specify and adjust planning decisions closer to the time of actual healthcare delivery, so that more detailed and accurate information can be incorporated [271, 329, 540]. As a result, it provides opportunities to better match care supply with fluctuating demand. By increasing the level of flexibility, an organization is able to on the one hand maintain a high level of delivery reliability by preventing that services cannot be delivered due to demand exceeding capacity. On the other hand, in periods of low demand, it is not burdened with surplus capacity that increases costs without a corresponding increase of revenues. Illustrations of flexibility reflected in this dissertation are those of care units sharing bed capacity when one of the units is fully occupied, and of deploying cross-trained nurses for who it is only at the start of a working shift decided in which care unit they will work.

In conclusion, the work in this thesis intends to make healthcare professionals even more aware of the added value of taking an integral perspective on logistical decision making. First, the problems addressed emphasize the importance of integrality in terms of objectives and performance: healthcare must be safe, effective, patient-centered, timely, efficient, and equitable. While the traditional belief is that quality and efficiency always confront each other, various examples strengthen our belief that they often can, and must, go hand in hand. Second, the research outcomes show the value of integrality in planning and control: performance is enhanced by aligning long-, medium-, and short-term decision making and by realizing coordination and collaboration between the various care chain actors. By consistently addressing the notions of variability and flexibility along the way, this dissertation aims to contribute the achievement of quality-driven efficiency in healthcare.

1.3 The role of Operations Research

The field of Operations Research and Management Science (OR/MS) is an interdisciplinary branch of applied mathematics, engineering and sciences that uses various scientific research-based principles, strategies, and analytical methods including mathematical modeling, statistics and algorithms to improve an organization's ability to enact rational and meaningful management decisions [324]. OR/MS has been widely applied to diverse areas such as manufacturing, telecommunications, transportation and service industries like airlines, hotel chains and retail. Since the 1950s, the application of OR/MS to healthcare has shown that it can also play a significant role in addressing the challenges healthcare faces. The last decade of the 20th century has shown an expansion in the breadth and volume of OR/MS applied to healthcare. Application areas include public policy [77, 653], performance analysis [393, 467], medical decision making [154, 483], and resource capacity planning and control [271, 272].

With respect to OR/MS that quantitatively supports and rationalizes decision making in resource capacity planning and control, many different topics have been addressed, such as operating room planning [99, 262], nurse staffing [91, 197] and appointment scheduling in outpatient clinics [104, 267]. In Chapter 2, this body of literature is structurally reviewed. Due to the interdisciplinary nature of OR/MS applied to healthcare, the extensive base of literature is published across various academic fields. To be better able to retrieve references from this broad availability, with the Center for Healthcare Operations Improvement and Research (CHOIR) of the University of Twente, we introduced and maintain the online literature database 'ORchestra' [319], in which references in the field of OR/MS in healthcare are categorized by medical and mathematical subject. All the articles mentioned in Chapter 2 are included and categorized in ORchestra.

The process of investigating a real-world problem of concern via OR/MS starts with carefully observing and formulating the problem, including gathering all relevant data [304]. Although the word 'problem' is standard terminology in OR/MS, it can also stand for 'evaluation of opportunities' [15]. The next step is to construct a mathematical model that attempts to abstract the essence of the real problem. This model should be a sufficiently precise representation capturing the essential features of the situation so that the solutions and conclusions obtained from the model are also valid for the real world [304]. The experiments conducted to verify whether this is the case are referred to as 'model validation'. Next, by quantitatively predicting the consequences of potential solutions, the goal is to inform and make recommendations to decision makers so that they are eventually able to make the best possible decisions. The final step is to come to implementation of a solution. Because implementation often requires people to do things differently, it often meets with resistance [369]. Although implementation is likely to be a managerial action rather than that of the operations researcher, successful implementation of results can, especially in healthcare, in our opinion only be achieved when researchers and practitioners work closely together. Therefore, we believe that involving users

throughout the modeling and experimentation process is essential. This is what we did in all applications described in this thesis.

The value of OR/MS is contained in both its process and its outcomes. The process of modeling typically leads to better understanding and recognition of a problem, due to the necessity of structuring and identifying the key-characteristics of the real-world situation [15]. The outcomes of OR/MS models make it possible to prospectively assess the consequences of various alternative interventions, without actually changing the system. Modeling is highly suitable in healthcare settings, since experimenting in practice may induce risks for patients and field experimenting makes it difficult to control all variables, takes more time, is more costly, and offers less statistical reliability [77, 339]. In addition, healthcare environments are generally politically charged due to the medical autonomy of clinicians. Especially in such environments quantifying the impact of potential solutions helps to let ratio predominate over emotion, so that fact-based rather than feeling-based decision making is realized [250, 369].

The developed models presented in this dissertation all intend to capture the inherent complexity of healthcare processes, so to be able to accurately analyze the relation between system configurations and system performance. Many OR/MS techniques exists, which each have there own specific benefits and limitations (see [15, 304, 550, 565, 637] for introductory books). With the purpose to provide the best decision support in each particular problem setting, a diversity of OR/MS techniques (often in combination) is applied in this thesis:

Computer simulation. Technique to imitate the operation of a real-world system as it evolves over time by developing a 'simulation model'. A simulation model usually takes the form of a set of assumptions about the operation of the system, expressed as mathematical or logical relations between the objects of interest in the system [383, 637].

Heuristics. Systematic methods to optimize problems by creating and/or iteratively improving candidate solutions. Heuristics are applied when exact approaches take too much computation time. They do not guarantee an optimal solution is found [1, 637].

Markov processes. Mathematical models for the random evolution of a system satisfying the so-called Markov property: given the present (state of stochastic process), the future (evolution of the process) is independent of the past (evolution of the process) [565, 638].

Mathematical programming. Optimization models consisting of an objective function, representing a reward to be maximized or a (penalty) cost to be minimized, and a set of constraints that circumscribe the decision variables [335, 469, 521].

Queueing theory. Mathematical methods to model and analyze congestion and delays at service facilities, by specifying the arrival and departure processes for each of the queues of a system [510, 638].

Stochastic Petri nets. Mathematical formalism providing a graphical language for modeling systems with interacting concurrent components [448, 480]. Petri nets

consist of places marked by tokens, and transitions moving these tokens. In stochastic Petri nets random firing delays are associated with transitions [417].

1.4 Academic Medical Center Amsterdam

The research described in this dissertation is for a substantial part motivated by challenges faced in the organization of patient care at the Academic Medical Center (AMC) in Amsterdam, the Netherlands. The AMC, founded in 1983 as a merge between the Wilhelmina Gasthuis and the Binnengasthuis, is one out of eight university hospitals in the Netherlands and is affiliated with the University of Amsterdam. Being a university hospital, the AMC has three principal tasks. Its primary task is patient care. In addition, the AMC carries out medical research and provides medical education [3]. The focus in patient care is to perform procedures known as top referral patient care. This is care associated with special, often expensive and complex, diagnostic procedures and treatment. Around 60% of the patients visit the AMC for top referral care. The service area for top referral patient care covers the whole of the Netherlands. The AMC also serves as a 'general hospital' for the population of the multi-cultural urban area surrounding the south-east of Amsterdam.

In 2011, the AMC had 1,002 registered beds, employed 7,041 people, and performed 30,129 clinical admissions, 31,086 day care admissions, and 387,549 outpatient visits [5]. In its current form, the AMC is organized in ten divisions, which are centrally supported by corporate staff and facility services. Like many Dutch hospitals the AMC faces rising demand, tight budget restrictions, and labor shortages [4]. In addition, the complexity of the provided care increases. To retain its position among the top medical centers in the world, the board of the AMC endorses the necessity of a fundamental reconsideration of the employed activities and a complete redesign of its operations.

The research described in this thesis has been performed in collaboration with the corporate staff department 'Quality Assurance and Process Innovation' (Kwaliteit en Procesinnovatie; KPI). Since 2008, the author of this dissertation has been a member of this department as a 'consultant process optimization'. The department KPI has the goal to support other AMC departments with monitoring and improving the quality of patient care. KPI employs a multidisciplinary team of consultants and connects consultancy with scientific research. It performs research on a broad area of quality improvement and patient safety. The research is carried out in close cooperation with other internal and external departments involved in improving patient care, patient logistics, patient centeredness, patient satisfaction, shared decision making, decision support techniques, evidence-based decision making, evidence-based practice, guideline adherence, management quality circles, safety management, quality indicators, clinical governance, medical & nursing audit and quality of care evaluation. This thesis is a result of a collaboration between KPI and the knowledge center CHOIR of the University of Twente.

As an academic medical center, the AMC chooses to apply scientific analysis tools and methodologies in redesigning patient care processes [191], with the under-

lying goal to not only deliver evidence-based patient care, but also to propagate knowledge-based based management. This is put into practice via the improvement program called 'SLIM' (referring to 'lean', and also meaning 'smart' in Dutch), in which the department KPI plays a leading role. SLIM is aimed at achieving increased levels of quality and efficiency in all primary and secondary services within the hospital. The work presented in this thesis connects with the goals formulated within the framework of SLIM.

The following focus areas of SLIM are specifically addressed in this thesis. With regards to outpatient care the AMC wants to encourage the possibilities of one-stop shopping and combination appointments, so that the number of outpatient visits per patient can be reduced. Other developments that are promoted are those of introducing more multidisciplinary care teams and providing automated support for appointment scheduling. Looking at inpatient care, a shift from clinical admissions to day care treatments is pursued, next to a reduction in the length of stays of clinical admissions, thereby reducing the number of required overnight stays. Then, by reducing the total number of beds, consolidating medical care units, and introducing flexible nurse pools, improvements in the efficient and effective use of beds and staff are strived for. Since the described developments and objectives are common to many present-day healthcare providers, and since our mathematical models are generically formulated, the application of the models and the relevance of their derived conclusions are not at all limited to the setting of the AMC.

1.5 Outline of this thesis

This thesis is organized in six parts. Part I is formed by this introductory chapter. Part II provides a general overview of the field of resource capacity planning and control in healthcare and a review of the state of the art in OR/MS. It sets up the conceptual framework within which several specific decision problems are studied in the following parts. Parts III-VI are organized according to the order of encounter in a typical patient's pathway. Part III focuses on combination appointments during single outpatient visits, Part IV on multidisciplinary treatments requiring a series of outpatient visits, Part V on inpatient care services, and Part VI on entire care pathways.

Part II comprises **Chapter 2** and provides a comprehensive overview of the typical decisions to be made in resource capacity planning and control in healthcare, in addition to a structured review of relevant OR/MS articles for each planning decision. Its contribution is twofold. First, to position the planning decisions, we present a taxonomy. This taxonomy provides healthcare managers and OR/MS researchers with a method to identify, break down and classify planning and control decisions. Second, following the taxonomy, for six healthcare services, we provide an exhaustive specification of resource capacity planning and control decisions. For each identified decision, we structurally review the key OR/MS articles and the OR/MS methods and techniques that are applied in the literature to support decision making.

Part III presents two studies that have the purpose to support the realization of one-stop shopping at ambulatory care services. In many settings it is highly valuable to patients to offer the combination of consultations, diagnostics, and treatments during a single visit. By one-stop shopping the number of hospital visits can be reduced, and required treatments can earlier be commenced and better be coordinated.

Chapter 3 is directed to outpatient clinics and diagnostic facilities that facilitate walk-in service, to improve accessibility, to offer more freedom for patients to choose their preferred time and date of visit, and to allow patients to visit multiple care providers on one day. The chapter shows the advantages of offering combined walk-in and scheduled service.

Chapter 4 provides an example of how OR/MS can support focused care facilities that offer multidisciplinary care to patients with specific complex diseases. The example concerns the 'Children's Muscle Center Amsterdam', which was opened in 2011 by the AMC to diagnose and treat children with neuromuscular diseases. Through the establishment of the center, clinical alignment is improved and children will generally visit the hospital only once a year instead of four to ten times.

Part IV is directed to rehabilitation care. Rehabilitation care is a treatment process that involves a series of treatments by therapists of various disciplines. These therapists may be affiliated with different departments and may use different planning horizons. This multidisciplinary nature of the rehabilitation process complicates planning and control. Improving coordination and alignment between different disciplines positively affects both quality and efficiency.

Chapter 5 presents a methodology to schedule treatments for rehabilitation outpatients entirely at once. This integral treatment planning methodology ensures continuity of the rehabilitation process while improving performance on various indicators among which access times, therapist utilization, and the ability to schedule combination appointments. The approach is applied to the rehabilitation outpatient clinic of the AMC.

Chapter 6 connects with the observation made at the end of Chapter 5, which states that balancing discipline capacities is a promising direction for further improvement. We perform an integral patient flow analysis for a case study of the rehabilitation center 'Het Roessingh', to support the implementation of treatment plans that are similar to those of Chapter 5.

Part V supports the design and operations of inpatient care services. Effectively designing inpatient care services requires simultaneous consideration of several interrelated planning issues, such as case mix, care unit partitioning, care unit size, and staffing decisions. The inpatient care facility is a downstream department of which the workload is mainly determined by the patient outflow of the operating theater and the emergency department. Therefore, coordination with surgical and emergency care services is essential. Workload on nursing wards depends highly on patient arrivals and patient lengths of stay, which are both inherently variable. Predicting this workload, and staffing nurses accordingly, is essential for guaranteeing quality of care in a cost effective manner.

Chapter 7 presents a model to predict bed census on nursing wards by hour as a function of the operating room schedule and a cyclic arrival pattern of emergency patients. The model enables the evaluation of alternative interventions with respect to both the design and the operations of inpatient care units. The effectiveness of the model is demonstrated by applying it to a case study of four surgical nursing wards of the AMC.

Chapter 8 introduces a method which takes the hourly census predictions from Chapter 7 as starting point to derive efficient nurse staffing policies. It particularly explores the potential of flexible staffing policies which allows hospitals to dynamically respond to their fluctuating patient population. The flexible policies involve the employment of so-called float nurses for whom it is only at the start of a working shift decided in which specific care units they will work. The method is applied to the same case study as that of Chapter 7.

Part VI intends to model entire patient care pathways. These pathways are generally stochastic and various patient flows share different resources. Typical questions arising when designing healthcare organizations are the identification of bottlenecks, achievable throughput and maximization of resource utilization. Therefore, performance analysis is an important issue in the design and implementation of healthcare systems. We believe that stochastic Petri nets are an appropriate formalism to model interacting care pathways in healthcare organizations. In these chapters, we build a theoretical foundation for a decision support tool along which we believe vital insight in the behavior of healthcare networks can be obtained.

Chapter 9 serves as an introduction to the chapters that follow by outlining elementary Petri nets definitions, properties, and results, and by providing a review of relevant literature.

Chapter 10 focuses on analytical (so-called product form) results, to create the conditions for efficient computation of relevant performance measures via closed-form expressions.

Chapters 11 and 12 formulate decomposition results that contribute to greater understanding of network behavior and performance, as they enable studying a system by analyzing the characteristics of separate components.

Chapter 13 takes the described results as starting point, to sketch directions for future research aimed at constructing and evaluating stochastic Petri nets based on patient event logs, thereby becoming able to deliver practical decision support.

The thesis closes with an epilogue, which summarizes our results and discusses the challenges encountered when implementing these.

Part II

A Taxonomy for Resource Capacity Planning and Control

CHAPTER 2

Structured Review of the State of the Art in Operations Research

2.1 Introduction

In Chapter 1, resource capacity planning and control in healthcare, and Operations Research and Management Sciences (OR/MS) were introduced and defined. In the current chapter, we provide a structured overview of the typical decisions to be made in resource capacity planning and control in healthcare, and we provide a review of the relevant OR/MS literature for each planning decision. First, a taxonomy is formulated to identify and position planning and control decisions. This taxonomy is the starting point to obtain a complete specification of planning decisions, and to gain understanding of the interrelations between various planning decisions. Herewith, we aim to guide healthcare professionals and OR/MS researchers through the broad field of OR/MS in healthcare. On the one hand, healthcare professionals can identify lacking, insufficiently defined and incoherent planning decisions within their department or organization. On the other hand, it gives the opportunity to identify decisions that are not yet addressed often in the OR/MS literature.

The contribution of this chapter is twofold. First, to position the planning decisions, we present a taxonomy. This taxonomy provides healthcare managers and OR/MS researchers with a method to identify, break down and classify planning and control decisions. The taxonomy contains two axes. The vertical axis reflects the hierarchical nature of decision making in resource capacity planning and control, and the horizontal axis the various healthcare services. The vertical axis is strongly connected, because higher-level decisions demarcate the scope of and impose restrictions on lower-level decisions. Although healthcare delivery is generally organized in autonomous organizations and departments, the horizontal axis is also strongly interrelated as a patient pathway often consists of several healthcare services from multiple organizations or departments.

Second, following the vertical axis of the taxonomy, and for each healthcare service on the horizontal axis, we provide a comprehensive specification of planning and control decisions in resource capacity planning and control. For each planning and control decision, we structurally review the key OR/MS articles and the OR/MS techniques that are applied in the literature to support decision making. No struc-

tured review exists of this nature, as existing reviews are typically exhaustive within a confined scope, such as simulation modeling in healthcare [339] or outpatient appointment scheduling [104], or are more general to the extent that they do not focus on the concrete specific decisions.

This chapter is organized as follows. Section 2.2 presents the taxonomy. Section 2.3 states the objectives, defines the scope, and summarizes the search method for the literature review. With the taxonomy as the foundation, Sections 2.4-2.6 identify, classify and discuss the planning and control decisions. Section 2.7 concludes this chapter with a discussion of our findings.

2.2 Taxonomy

Taxonomy is the practice and science of classification. It originates from biology where it refers to a hierarchical classification of organisms. The National Biological Information Infrastructure [452] provides the following definition of taxonomy: "Taxonomy is the science of classification according to a pre-determined system, with the resulting catalog used to provide a conceptual framework for discussion, analysis, or information retrieval; ...a good taxonomy should be simple, easy to remember, and easy to use." With exactly these objectives, we present a taxonomy for resource capacity planning and control in healthcare.

Planning and control decisions are made by healthcare organizations to design and operate the healthcare delivery process. It requires coordinated long-term, medium-term and short-term decision making in multiple managerial areas. In [273], a framework is presented to subdivide these decisions in four hierarchical, or temporal, levels and four managerial areas. These hierarchical levels and the managerial area of resource capacity planning and control form the basis for our taxonomy. For the hierarchical levels, [273] applies the well-known breakdown of *strategic*, *tactical* and *operational* [17]. In addition, the operational level is subdivided in *offline* and *online* decision making, where *offline* reflects the in advance decision making and *online* the real-time reactive decision making in response to events that cannot be planned in advance. The four managerial areas are: medical planning, financial planning, materials planning and resource capacity planning. They are defined as follows. *Medical planning* comprises decision making by clinicians regarding medical protocols, treatments, diagnoses and triage. *Financial planning* addresses how an organization should manage its costs and revenues to achieve its objectives under current and future organizational and economic circumstances. *Materials planning* addresses the acquisition, storage, distribution and retrieval of all consumable resources/materials, such as suture materials, blood, bandages, food, etc. *Resource capacity planning* addresses the dimensioning, planning, scheduling, monitoring, and control of renewable resources. Our taxonomy is a refinement of the healthcare planning and control framework of [273] in the resource capacity planning area.

The taxonomy contains two axes. The vertical axis reflects the hierarchical nature of decision making in resource capacity planning and control, and is derived

from [273]. On the horizontal axis of our taxonomy we position different services in healthcare. We identify *ambulatory care services*, *emergency care services*, *surgical care services*, *inpatient care services*, *home care services*, and *residential care services*. The taxonomy is displayed in Figure 2.1. We elaborate on both axes in detail below.

Vertical axis

Our taxonomy is intended for planning and control decisions within the boundaries of a healthcare delivery organization. Every healthcare organization operates in a particular external environment. Therefore, all planning and control decisions are made in the context of this external environment. The external environment is characterized by factors such as legislation, technology and social factors.

The nature of planning and control decision making is such that decisions disaggregate as time progresses and more information becomes available [654]. Aggregate decisions are made in an early stage, while more detailed information supports decision making with a finer granularity in later stages. Because of this disaggregating nature, most well-known taxonomies and frameworks for planning and control are organized hierarchically [273, 654]. As the impact of decisions decreases when the level of detail increases, such a hierarchy also reflects the top-down management structure of most organizations [51].

For completeness we explicitly state the definitions of the four hierarchical levels of [273], which we position on the vertical axis of our taxonomy. The definitions are adapted to specifically fit the managerial area of resource capacity planning and control.

Strategic planning addresses structural decision making. It involves defining the organization's mission (i.e., 'strategy' or 'direction'), and the decision making to translate this mission into the design, dimensioning, and development of the healthcare delivery process. Inherently, strategic planning has a long planning horizon and is based on highly aggregated information and forecasts. Examples of strategic planning are determining facility locations, dimensioning resource capacities (e.g., acquisition of an MRI scanner, staff) and deciding on the service and case mix.

Tactical planning translates strategic planning decisions to guidelines which facilitate operational planning decisions. While strategic planning addresses structural decision making, tactical planning addresses the organization of the operations/execution of the healthcare delivery process (i.e., the 'what, where, how, when and who'). As a first step in tactical planning, patient groups are characterized based on disease type/diagnose, urgency and resource requirements. As a second step, the available resource capacities, settled at the strategic level, are divided among these patient groups. In addition to the allocation in time quantities, more specific timing information can already be added, such as dates or time slots. In this way, blueprints for the operational planning are created that allocate resources to different tasks, specialties and patient groups. Temporary capacity expansions like overtime or hiring staff are also part of tactical

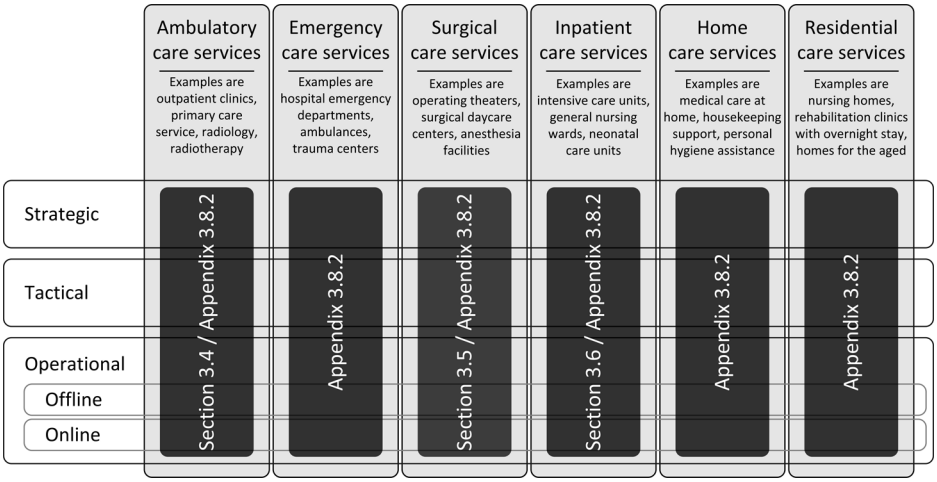


Figure 2.1: The taxonomy for resource capacity planning and control decisions in healthcare.

planning. Demand has to be (partly) forecasted, based on (seasonal) demand, waiting list information, and the 'downstream' demand in care pathways of patients currently under treatment. Examples of tactical planning are staff-shift scheduling and the (cyclic) surgical block schedule that allocates operating time capacity to patient groups.

Operational planning (both 'offline' and 'online') involves the short-term decision making related to the execution of the healthcare delivery process. Following the tactical blueprints, execution plans are designed at the individual patient level and the individual resource level. In operational planning, elective demand is entirely known and only emergency demand has to be forecasted. In general, the capacity planning flexibility is low on this level, since decisions on higher levels have demarcated the scope for the operational level decision making.

Offline operational planning reflects the in advance planning of operations. It comprises the detailed coordination of the activities regarding current (elective) demand. Examples of offline operational planning are patient-to-appointment assignment, staff-to-shift assignment and surgical case scheduling.

Online operational planning reflects the control mechanisms that deal with monitoring the process and reacting to unplanned events. This is required due to the inherent uncertain nature of healthcare processes. An example of online operational planning is the real-time dynamic (re)scheduling of elective patients when an emergency patient requires immediate attention.

Note that the decision horizon lengths are not explicitly defined for any of the hierarchical planning levels, since these depend on the specific characteristics of the application. For example, an emergency department inherently has shorter planning horizons than a long-stay ward in a nursing home. Furthermore, there is a strong

interrelation between hierarchical levels. Top-down interaction exists as higher-level decisions demarcate the scope of and impose restrictions on lower-level decisions. Conversely, bottom-up interaction exists as feedback about the healthcare delivery realization supports decision making in higher levels.

Horizontal axis

On the horizontal axis of our taxonomy we position the different services in healthcare. The complete spectrum of healthcare delivery is a composition of many different services provided by many different organizations. From the perspective of resource capacity planning and control, different services may face similar questions. To capture this similarity, we distinguish six clusters of healthcare services. The definitions of the six care services are obtained from the corresponding MeSH terms provided by PubMed [574]. For each care service we offer several examples of facilities that provide this service.

Ambulatory care services provide care to patients without offering a room, a bed and board, and they may be free-standing or part of a hospital. Examples of ambulatory care facilities are outpatient clinics, primary care services and the hospital departments of endoscopy, radiology and radiotherapy.

Emergency care services are concerned with the evaluation and initial treatment of urgent and emergent medical problems, such as those caused by accidents, trauma, sudden illness, poisoning, or disasters. Emergency medical care can be provided at the hospital or at sites outside the medical facility. Examples of emergency care facilities are hospital emergency departments, ambulances and trauma centers.

Surgical care services provide operative procedures (surgeries) for the correction of deformities and defects, repair of injuries, and diagnosis and cure of certain diseases. Examples of surgical care facilities are the hospital's operating theater, surgical daycare centers and anesthesia facilities.

Inpatient care services provide care to hospitalized patients by offering a room, a bed and board. Examples are intensive care units, general nursing wards, and neonatal care units.

Home care services are community health and nursing services that provide multiple, coordinated services to a patient at the patient's home. Home care services are provided by a visiting nurse, home health agencies, hospitals, or organized community groups using professional staff for healthcare delivery. Examples are medical care at home, housekeeping support and personal hygiene assistance.

Residential care services provide supervision and assistance in activities of daily living with medical and nursing services when required. Examples are nursing homes, psychiatric hospitals, rehabilitation clinics with overnight stay, homes for the aged, and hospices.

Note that the horizontal subdivision is not based on healthcare organizations, but on the provided care services. Therefore, it is possible that a single healthcare organization offers services in multiple clusters. It may be that a particular facility is used by multiple care services, for example a diagnostics department that is used in both ambulatory and emergency care services. In addition, a patient's treatment often comprises of consecutive care stages offered by multiple care services. The healthcare delivery realization within one care service is impacted by decisions in other services, as inflow and throughput strongly depend on these other services. Therefore, resource capacity planning and control decisions are always made in the context of decisions made for other care services. Hence, like the interrelation in the vertical levels, a strong interrelation exists between the horizontal clusters.

2.3 Objectives, scope, and search method

In this section, with our taxonomy as the foundation, we provide an exhaustive specification of planning decisions in healthcare, combined with a review of key OR/MS literature. We identify the resource capacity planning and control decisions for for each of the six care services in our taxonomy. The decisions are classified according to the vertical hierarchical structure of our taxonomy. For each identified planning decision we will discuss the following in our overview:

- What is the concrete *decision*?
- Which *performance measures* are considered?
- What are the *key trade-offs*?
- What are *main insights and results* from the literature?
- What are *general conclusions*?
- Which *OR/MS methods* are applied to support decision making?

The identified planning decisions are in the first place obtained from available books and articles on healthcare planning and control. Our literature search method is explained in more detail below. In addition, to be as complete as possible, expert opinions from healthcare professionals and OR/MS specialists are obtained to identify decisions that are not yet well-addressed in the literature and for this reason cannot be obtained from the literature. In the rest of this section, we discuss the scope of the identified planning decisions and the applied OR/MS methods, and present the applied literature search method.

Scope. Numerous processes are involved in healthcare delivery. We focus on the resource capacity planning and control decisions to be made regarding the *primary process* of healthcare delivery. In the management literature, the primary process is defined as the set of activities that are directly concerned with the creation or delivery of a product or service [485]. Thus, we do not focus on *supporting activities*, such as procurement, information technology, human resource management, laboratory services, blood services and instrument sterilization.

We focus on OR/MS methods that quantitatively support and rationalize decision making in resource capacity planning and control. Based on forecasting of demand for care (see [468] for forecasting techniques), these methods provide optimization techniques for the design of the healthcare delivery process. Outside our scope is statistical comparison of performance of healthcare delivery organizations, so-called benchmarking, of which Data Envelopment Analysis (DEA) and Stochastic Frontier Analysis (SFA) are well-known examples [127]. Quantitative decision making requires measurable performance indicators by which the quality of healthcare delivery can be expressed. A comprehensive survey of applied performance measures in healthcare organizations is provided in [393]. Next, practical implementation of OR/MS methods may require the development of ICT solutions (that are possibly integrated in healthcare organizations' database systems); this is also outside the scope of our review.

The spectrum of different OR/MS methods is wide (see for example [304, 550, 565, 637] for introductory books). In this review, we distinguish the following OR/MS methods: computer simulation [383], heuristics [1], Markov processes (including Markov reward and decision processes) [565], mathematical programming [469, 521], queueing theory [510]. In Chapter 1, a short description of each of these OR/MS methods was provided.

Literature search method. As the body of literature on resource capacity planning and control in healthcare is extensive, we applied a structured search method in which we restricted ourselves to articles published in ISI-listed journals to ensure that we would find and filter key and state-of-the-art contributions. Figure 2.2 displays our search method. To identify the search terms as listed in Appendix 2.8.1 and to create the basic structure of the planning decision hierarchy for each care service, we consulted available literature reviews [58, 74, 76, 91, 99, 104, 118, 197, 218, 219, 262, 266, 267, 328, 339, 344, 346, 369, 405, 428, 441, 450, 475, 483, 488, 495, 499, 540, 541, 571, 591] and books [77, 271, 378, 437, 468, 608]. Additional search terms were obtained from the index of *Medical Subject Headings* (MeSH) [574] and available synonyms. With these search terms, we performed a search on the database of Web of Science (WoS) [564]. We chose WoS as it contains articles from all ISI-listed journals. It is particularly useful as it provides the possibility to select *Operations Research and Management Science* as a specific subject category and to sort references on the number of citations.

We identified a base set containing the ten most-cited articles in the predefined subject category of *Operations Research and Management Science*. Starting from this base set, we included all articles from ISI-listed journals that are referred by or refer to one of the articles in the base set and deal with resource capacity planning and control decisions. As such, we ensured that we also reviewed recent work that may not have been cited often yet. In addition, we included articles published in Health Care Management Science (HCMS), which is particularly relevant for OR/MS in healthcare and obtained an ISI listing in 2010. To be sure that by restricting to WoS and HCMS, we did not neglect essential references, we also performed a search with our search terms on the databases of Business Source Elite [188], PubMed [575]

Step 1:	Identify search terms from reviews, books and MeSH
Step 2:	Search the OR/MS subject category in WoS with the search terms
Step 3:	Select a base set: the ten most-cited articles relevant for our review
Step 4:	Perform a backward and forward search on the base set articles
Step 5:	Search relevant articles from HCMS

Figure 2.2: The search method applied to each care service.

and Scopus [194]. This search did not result in significant additions to the already found set of papers. The literature search was updated up to May 10, 2012.

In the following sections, we provide a selection of the reviews per care service. Section 2.4 is devoted to ambulatory care services, Section 2.5 to surgical care services, and Section 2.6 to inpatient care services. For the reviews of emergency care services, home care services, and residential care services, we refer the reader to [320]. For each care service, the review is subdivided in strategic, tactical, offline operational and online operational planning. In Appendix 2.8.2, we do present tables for all six care services in which the identified planning decisions are listed, together with applied OR/MS methods and literature references per planning decision. When for different care services a similar planning decision is involved, we use the same term. It is our intention that Sections 2.4-2.6 are self-contained, so that they can be read in isolation. Therefore, minor passages are overlapping. When in the description of a planning decision a paper is cited, while it does not appear in the 'methods'-list, it means that this paper contains a relevant statement about this planning decision, but the particular planning decision is not the main focus of the paper.

2.4 Ambulatory care services

Ambulatory care services provide medical interventions without overnight stay, i.e., the patient arrives at the facility and leaves the facility on the same day. These medical interventions comprise for example diagnostic services (e.g, CT scans, MRI scans), doctor consultations, radiotherapy treatments or minor surgical interventions. Demand for ambulatory care services is growing in most western countries since 2000 [466]. The existing literature has mainly focused on the offline operational planning decision of appointment scheduling.

Strategic planning

Regional coverage. Ambulatory care planning on a regional level aims to create the infrastructure to provide healthcare to the population in its catchment area. This regional coverage decision involves determining the number, size and location of facilities in a certain region to find a balanced distribution of facilities with respect

to the geographical location of demand [181]. The main trade-off in this decision is between patient accessibility and efficiency. Patient accessibility is represented by access time and travel distance indicators. Efficiency is represented by utilization and productivity indicators [181, 540]. Common regional planning models incorporate the dependency of demand on the regional demographic and socioeconomic characteristics [2].

Methods: computer simulation [425, 505, 543, 562], heuristics [2, 181], literature review [540].

Service mix. An organization decides the particular services that the ambulatory care facility provides. The service mix stipulates which patient types can be consulted. In general, the service mix decision is not made at an ambulatory care service level, but at the regional or hospital level, as it integrally impacts the ambulatory, emergency, surgical and inpatient care services. This is also expressed by [606] in which for example inpatient resources, such as beds and nursing staff, are indicated as 'following' resources. This may be the reason that we have not found any papers focusing on service mix decisions for ambulatory care services in specific.

Methods: no papers found.

Case mix. Every ambulatory care facility decides on a particular case mix, which is the volume and composition of patient groups that the facility serves. The settled service mix restricts the decisions to serve particular patient groups. Patient groups can be classified based on disease type, age, acuteness, home address, etc. The case mix influences almost all other planning decisions, such as a facility's location, capacity dimensions and layout. Also, demand for different patient groups in the case mix may vary, which influences required staffing levels significantly [539, 548]. However, case mix decision making has not received much attention in the OR/MS literature. In the literature, the case mix is often treated as given.

Methods: computer simulation [548], mathematical programming [539].

Panel size. The panel size is the number of potential patients of an ambulatory care facility [256]. Since only a fraction of these potential patients, also called calling population, actually demands healthcare, the panel size can be larger than the number of patients a facility can serve. The panel size is particularly important for general practitioners, as they need an accurate approximation of how many patients they can subscribe or admit to their practice. A panel size should be large enough to have enough demand to be profitable and to benefit from economies of scale, as a facility's costs per patient decrease when the panel size increases [543]. On the other hand, when the panel size is too large, access times may grow exponentially [256].

Methods: computer simulation [543], queueing theory [256].

Capacity dimensioning. Ambulatory care facilities dimension their resources, such as staff, equipment and space, with the objective to (simultaneously) maximize clinic profit, patient satisfaction, and staff satisfaction [548]. To this end, provider capacity must be matched with patient demand, such that performance measures such as costs, access time and waiting time are controlled. Capacity is dimensioned for the following resource types:

- *Consultation rooms*. The number of consultation rooms that balances patient waiting times and doctor idle time with costs for consultation rooms [321, 540, 547, 548].
- *Staff*. Staff in the ambulatory care services concern for example doctors, nurses and assistants [43, 339, 425, 506, 539, 540, 543, 547, 548, 619, 626].
- *Consultation time capacity*. The total consultation time that is available, for example for an MRI scanner or a doctor [139, 189, 193].
- *Equipment*. Some ambulatory care services require equipment for particular consultation types, for example MRI scanners, CT scanners and radiotherapy machines [225, 425, 562].
- *Waiting room*. The waiting room is dimensioned such that patients and their companions waiting for consultation can be accommodated [548].

When capacity is dimensioned to cover average demand, variations in demand may cause long access and waiting times [562]. Basic rules from queueing theory demonstrate the necessity of excess capacity to cope with uncertain demand [251]. Capacity dimensioning is a key decision, as it influences how well a facility can meet demand and manage access and waiting times.

Methods: computer simulation [189, 193, 225, 321, 425, 506, 543, 547, 548, 562, 626], Markov processes [619], mathematical programming [539], queueing theory [43, 139, 193, 321], literature review [339, 540].

Facility layout. The facility layout concerns the positioning and organization of various physical areas in a facility. A typical ambulatory care facility consists of a reception area, a waiting area, and consultation rooms [228]. The facility layout is a potentially cost-saving decision in ambulatory care facilities [228, 468], but we found no papers that used an OR/MS approach to study the layout of an ambulatory care facility. Yet, the handbook [468] discusses heuristics for facility layout problems in healthcare.

Methods: heuristics [468].

Tactical planning

Patient routing. Ambulatory care typically consists of multiple stages. We denote the composition and sequence of these stages as the route of a patient. An effective and efficient patient route should match medical and capacity requirements, and the facility's layout. For a single facility, identifying different patient types and designing customized patient routes for each type prevents superfluous stages and delays [425]. For example, instead of two visits to a doctor and a medical test in between, some patient types may undergo a medical test before visiting the doctor, which saves valuable doctor time. Parallel processing of patients may increase utilization of scarce resources (e.g., a doctor or a CT scanner) [225, 321]. When parallel processing is applied, idle time of the scarce resource is reduced by preparing patients for consultation during the consultation time of other patients. Performance is typically measured by total visit time, waiting time, and queue length.

Methods: computer simulation [113, 225, 321, 425, 543], queueing theory [321, 655], literature review.

Capacity allocation. On the tactical level, resource capacities settled on the strategic level are subdivided over all patient groups. To do so, patient groups are first assigned to resource types.

- *Assign patient groups to resource types.* The assignment of patient groups to available resources requires knowledge about the capabilities of for example clinical staff, support staff or medical equipment, and the medical characteristics of patients. The objective is to maximize the number of patients served, by calculating the optimal assignment of patient groups to appropriately skilled members of clinical staff [539]. Efficiency gains are possible when certain tasks can be substituted between clinical staff, either horizontally (equally skilled staff) or vertically (lower skilled staff) [540].
- *Time subdivision.* The available resource capacities, such as staff and equipment, is subdivided over patient groups. For example, general practitioners divide their time between consulting patients and performing prevention activities for patients [265]. When patient demand changes over time (e.g., seasonality), a dynamic subdivision of capacity, updated based on current waiting lists, already planned appointments and expected requests for appointments, performs better than a long-term, static subdivision of resource capacity [604].

Methods: computer simulation [604], mathematical programming [265, 539], literature review [608].

Temporary capacity change. The balance between access times and resource utilization may be improved when resource capacities can temporarily be increased or decreased, to cope with fluctuations in patient demand [604]. For example, changing a CT scanner's opening hours [604] or changing doctor consultation time [193].

Methods: computer simulation [193, 604].

Access policy. In appointment-driven facilities, the access policy concerns the waiting list management that deals with prioritizing waiting lists so that access time is equitably distributed over patient groups. In the traditional approach, there is one queue for each doctor, but when patient queues are pooled into one joint queue, patients can be treated by the first available doctor, which reduces access times [599]. Another policy is to treat patients without a scheduled appointment, also called 'walk-in' service. In between scheduled and walk-in service is 'advanced access' (also called 'open access', or 'same-day scheduling'). With advanced access, a facility leaves a fraction of the appointment slots vacant for patients that request an appointment on the same day or within a couple of days. The logistical difficulty of both walk-in service and advanced access is a greater risk of resource idle time, since patient arrivals are more uncertain. However, implementation of walk-in/advanced access can provide significant benefits to patient access time, doctor idle time and doctor overtime, when the probability of patients not showing up is relatively large [472, 504]. A proper balance between traditional appointment

planning and walk-in/advanced access further decreases access times and increases utilization [502, 655]. The specification of such a balanced design is a tactical planning decision, which will be discussed later in this section.

Methods: computer simulation [20, 212, 472, 502, 599], Markov processes [472], queueing theory [504, 655].

Admission control. Given the access policy decisions, admission control involves the rules according to which patients are selected to be admitted from the waiting lists. Factors that are taken into account are for example resource availability, current waiting lists and expected demand. Clearly, this makes admission control and capacity allocation mutually dependent. This is for example the case in [604], where the capacity subdivision for a CT scanner is settled by determining the number of patients to admit of each patient group. Access times can be controlled by adequate admission control [232, 237, 337, 604]. Admission control plays a significant role in advanced access or walk-in policies. Successful implementation of these policies requires a balance between the reserved and demanded number of slots for advanced access or walk-in patients. Too many reserved slots results in resource idle time, and too little reserved slots results in increased access time [490, 492].

Methods: computer simulation [604], heuristics [237], Markov processes [232, 237], mathematical programming [337, 490, 492].

Appointment scheduling. Appointment schedules are blueprints that can be used to provide a specific time and date for patient consultation (e.g., an MRI scan or a doctor visit). Appointment scheduling comprises the design of such appointment schedules. Typical objectives of this design are to minimize patient waiting time, maximize resource utilization or minimize resource overtime. A key trade-off in appointment scheduling is the balance between patient waiting time and resource idle time [104, 308, 345]. Appointment scheduling is comprehensively reviewed in [104, 267]. In an early paper [624], the Bailey-Welch appointment scheduling rule is presented, which is a robust and well-performing rule in many settings [308, 340, 357]. References differ in the extent in which various aspects are incorporated in the applied models. Frequently modeled aspects that influence the performance of an appointment schedule are patient punctuality [212, 390, 629], patients not showing up ('no-shows') [212, 213, 309, 340], walk-in patients or urgent patients [20, 212, 502, 655], doctor lateness at the start of a consultation session [212, 213, 400, 506], doctor interruptions (e.g., by comfort breaks or administration) [213, 390], and the variance of consultation duration [308]. These factors can be taken into account when modeling the following key decisions that together design an appointment schedule.

- *Number of patients per consultation session.* The number of patients per consultation session is chosen to control patient access times and patient waiting times. When the number of patients is increased, access times may decrease, but patient waiting times and provider overtime tend to increase [100, 212, 308].
- *Patient overbooking.* Patients not showing up, also called 'no-shows', cause unexpected gaps, and thereby increase resource idleness [308]. Overbooking of

patients, i.e., booking more patients into a consultation session than the number of planned slots, is suggested to compensate no-shows in [363, 368, 386, 451, 542]. Overbooking can significantly improve patient access times and provider productivity, but it may also increase patient waiting time and staff overtime [363, 368]. Overbooking particularly provides benefits for large facilities with high no-show rates [363].

- *Length of the appointment interval.* The decision for the length of the planned appointment interval or slot affects resource utilization and patient waiting times. When the slot length is decreased, resource idle time decreases, but patient waiting time increases [213]. For some distributions of consultation time, patient waiting times and resource idle time are balanced when the slot length equals the expected length of a consultation [104]. The slot length can be chosen equal for all patients [213, 308, 624], but using different, appropriate slot lengths for each patient group may decrease patient waiting time and resource idle time when expected consultation times differ between patient groups [189].
- *Number of patients per appointment slot.* Around 1960, it was common to schedule all patients in the first appointment slot of a consultation session [220]. This minimizes resource idle time, but has a negative effect on patient waiting times [483, 506]. Later, it became common to distribute patients evenly over the consultation session to balance resource idle time and patient waiting time. In [220] various approaches in between these two extremes are evaluated, such as two patients in one time slot and zero in the next.
- *Sequence of appointments.* When different patient groups are to be scheduled, the sequencing of appointments influences waiting times and resource utilization. Appointments can be sequenced based on patient group or expected variance of the appointment duration. In [357] various rules for patient sequencing are compared. Alternatively, when differences between patients exist with respect to the variation of consultation duration, sequencing patients by increasing variance (i.e., lowest variance first) may minimize patient waiting time and resource idle time [104].
- *Queue discipline in the waiting room.* The queue discipline in the waiting room affects patient waiting time, and the higher a patient's priority, the lower the patient's waiting time. The queue discipline in the waiting room is often assumed to be First-Come First-Served (FCFS), but when emergency patients and walk-in patients are involved, the highest priority is typically given to emergency patients and the lowest priority to walk-in patients [104]. Priority can also be given to the patient that has to visit the most facilities on the same day [425].
- *Anticipation for unscheduled patients.* Facilities that do also serve unscheduled patients, such as walk-in and urgent patients, require an appointment scheduling approach that anticipates these unscheduled patients by reserving slack capacity. This can be achieved by leaving certain appointment slots vacant [179], or by increasing the length of the appointment interval [104]. Reserving too little capacity for unscheduled patients results in an overcrowded facility, while

reserving too many may result in resource idle time. Often, unscheduled patients arrive in varying volumes during the day and during the week. When an appropriate number of slots is reserved for unscheduled patients, and appointments are scheduled at moments that the expected unscheduled demand is low, patient waiting times decrease and resource utilization increases [463, 502, 655]. In the online operational level of this section, we discuss referring unscheduled patients to a future appointment slot when the facility is overcrowded.

Methods: computer simulation [22, 100, 105, 160, 189, 212, 213, 278, 308, 309, 345, 368, 390, 400, 425, 463, 502, 548, 608, 609, 624, 629], heuristics [100, 340], Markov processes [220, 257, 340, 357, 396, 451, 542], mathematical programming [100, 155, 503], queueing theory [73, 139, 179, 363, 386, 503, 608, 655], literature review [104, 267, 339, 540].

Staff-shift scheduling. Shifts are hospital duties with a start and end time [91]. Shift scheduling deals with the problem of selecting what shifts are to be worked and how many employees should be assigned to each shift to meet patient demand [197]. More attractive schedules promote job satisfaction, increase productivity, and reduce turnover. While staff dimensioning on the strategic level has received much attention, shift scheduling in ambulatory care facilities seems underexposed in the literature. In [88], shift schedules are developed for physicians, who often have disproportionate leverage to negotiate employment terms, because of their specialized skills. Hence, physicians often have individual arrangements that vary by region, governing authority, seniority, specialty and training. Although these individual arrangements impose requirements to the shift schedules, there is often flexibility for shifts of different lengths and different starting times to cope with varying demand during the day or during a week. In this context, the handbook [468] discusses staggered shift scheduling and flexible shift scheduling. In the first alternative, employees have varying start and end times of a shift, but always work a fixed number of hours per week. In the latter, cheaper alternative, a core level of staff is augmented with daily adjustments to meet patient demand.

Methods: computer simulation [478], mathematical programming [88], literature review [91, 197, 271, 468].

Offline operational planning

Patient-to-appointment assignment. Based on the appointment scheduling blueprint developed on the tactical level, patient scheduling comprises scheduling of an appointment in a particular time slot for a particular patient. A patient may require multiple appointments on one or more days. Therefore, we distinguish scheduling a *single appointment*, *combination appointments* and *appointment series*.

- *Single appointment.* Patients requiring an appointment often have a preference for certain slots. When information is known about expected future appointment requests and the expected preferences of these requests, a slot can be planned for this patient to accommodate the current patient, but also to have sufficient slots available for future requests from other patients. This can for example be

necessary to ensure that a sufficient number of slots is available for advanced access patients [268, 621], or to achieve equitable access for all patient groups to a diagnostic facility [474].

- *Combination appointments.* Combination appointments imply that multiple appointments for a single patient are planned on the same day, so that a patient requires fewer hospital visits. This is the case when a patient has to undergo various radiotherapy operations on different machines within one day [481].
- *Appointment series.* For some patients, a treatment consisting of multiple (recurring) appointments may span a period of several weeks or months. The treatment is planned in an appointment series, in which appointments may have precedence relations and certain requirements for the time intervals in between. In addition, the involvement of multiple resources may further complicate the planning of the appointment series. The appointment series have to fit in the existing appointment schedules, which are partly filled with already scheduled appointments. Examples of patients that require appointment series are radiotherapy patients [132, 133, 135] and rehabilitation patients [120].

Methods: heuristics [120, 481, 621], Markov processes [268, 474, 621], mathematical programming [132, 133, 135].

Staff-to-shift assignment. On the tactical level, staff shift scheduling results in shifts that have to be worked. In staff-to-shift assignment on the offline operational level, a date and time are given to staff members to perform particular shifts. For example, a consultation session is scheduled for a doctor on a particular day and time, and with a certain duration. For an endoscopy unit, the authors of [337] develop a model to schedule available doctors to endoscopy unit shifts.

Methods: mathematical programming [337], literature review [271].

Online operational planning

Dynamic patient (re)assignment. After patients are assigned to slots in the appointment schedule, the appointments are carried out on their planned day. During such a day, unplanned events, such as emergency or walk-in patients, extended consultation times, and equipment breakdown, may disturb the planned appointment schedule. In such cases, real-time dynamic (re)scheduling of patients is required to improve patient waiting times and resource utilization in response to acute events. For example, to cope with an overcrowded facility walk-in patients can be rescheduled to a future appointment slot to improve the balance of resource utilization over time [497]. Dynamic patient (re)assignment can also be used to decide which patient group to serve in the next time slot in the appointment schedule [257], for example based on the patient groups' queue lengths. When inpatients are involved in such decisions, they are often subject to rescheduling [104], since it is assumed that they are less harmed by a rescheduled appointment as they are already in the hospital. However, longer waiting times of inpatients may be more costly, since it may mean they have to be hospitalized longer [146].

Methods: computer simulation [497], Markov processes [146, 257, 396], mathematical programming [146].

Staff rescheduling. At the start of a shift, the staff schedule is reconsidered. Before and during the shift, the staff capacities may be adjusted to unpredicted demand fluctuations and staff absenteeism by using part-time, on-call nurses, staff overtime, and voluntary absenteeism [261, 483].

Methods: no papers found.

2.5 Surgical care services

Surgeries are physical interventions on tissues, generally involving cutting of a patient's tissues or closure of a previously sustained wound, to investigate or treat a patient's pathological condition. Surgical care services have a large impact on the operations of the hospital as a whole [40, 59, 99], and they are the hospital's largest revenue center [99, 157]. Surgical care services include ambulatory surgical wards, where patients wait and stay before and after being operated. We do not classify such wards as inpatient care services, since patients served on ambulatory basis do not require an overnight stay. The proportion of ambulatory surgeries, which are typically shorter, less complex and less variable [482], is increasing in many hospitals [428]. There is a vast amount of literature on OR/MS in surgical care services, comprehensively surveyed in [58, 99, 161, 262, 266, 267, 405, 428, 488, 541, 613]. These surveys are used to create the *taxonomic* overview of the planning decisions.

Strategic planning

Regional coverage. At a regional level, the number, types and locations of surgical care facilities have to be determined to find a balanced distribution of facilities with respect to the geographical location of demand [181]. The main trade-off in this decision is between patient accessibility and facility efficiency. Coordination of activities between hospitals in one region, can provide significant cost reductions at surgical care facilities and downstream facilities [71, 513].

Methods: computer simulation [71], mathematical programming [513].

Service mix. An organization selects the particular services that the surgical care facility provides. The service mix stipulates which surgery types can be performed, and therefore impacts the net contribution of a facility [312]. Specific examples of services are medical devices to perform noninvasive surgeries and robotic services for assisting in specialized surgery [156]. In general, the service mix decision is not made at a surgical care service level, but at the regional or hospital level, as it integrally impacts the ambulatory, emergency, surgical and inpatient care services.

Methods: no papers found.

Case mix. The case mix involves the number and types of surgical cases that are performed at the facility. Often, diagnosis-related groups (DRGs), which classify patient

groups by relating common characteristics such as diagnosis, treatment and age to resource requirements, are used to identify the patient types included in the case mix [317]. The case mix is chosen with the objective to optimize net contribution while considering several internal and external factors [262, 317]. Internal factors include the limited resource capacity, the settled service mix, research focus, and medical staff preferences and skills [59, 262, 338]. External factors include societal preferences, the disease processes affecting the population in the facility's catchment area [59], the case mix of competing hospitals [177], and the restricted budgets and service agreements in government funded systems [59]. High profit patient types may be used to cross-subsidize the unprofitable ones, possibly included for research or societal reasons [59].

Methods: computer simulation [338], mathematical programming [59, 317], literature review [262].

Capacity dimensioning. Surgical care facilities dimension their resources with the objective to optimize hospital profit, idle time costs, surgery delays, access times, and staff overtime [403, 520]. Therefore, provider capacity must be matched with patient demand [520] for all surgical resource types. The capacity dimensioning decisions for different resource types are highly interrelated and performance is improved when these decisions are coordinated both within the surgical care facility and with capacity dimension decisions in services outside the surgical care facility, such as medical care units and the Intensive Care Unit (ICU) [77, 519, 590, 591]. The following resources are dimensioned:

- *Operating rooms*. Operating rooms can be specified by the type of procedures that can be performed [32, 266, 339, 520].
- *Operating time capacity*. This concerns the number of hours per time period the surgical care services are provided [338, 428, 519, 560, 590]. Operating time capacity is determined by the number of operating rooms and their opening hours [403].
- *Presurgical rooms*. These rooms are used for preoperative activities, for example induction rooms for anesthetic purposes [428].
- *Recovery wards*. At these wards, patients recover from surgery [365, 366, 367, 519, 520]. The recovery ward is also called the Post Anaesthesia Care Unit (PACU) [262].
- *Ambulatory surgical ward*. At this ward, outpatients stay before and after surgery.
- *Equipment*. Equipment may be required to perform particular surgeries. Examples are imaging equipment [267] or robotic equipment [156]. Equipment may be transferable between rooms, which increases scheduling flexibility.
- *Staff*. Staff in surgical care services include surgeons, anesthesiologists, surgical assistants and nurse anesthetists [9, 89, 156, 312]. Staffing costs are a large portion of costs in surgical care services [18, 156]. Significant cost savings can be achieved by increasing staffing flexibility [156], for example by (i) cross-training surgical assistants for multiple types of surgeries [266], (ii) augmenting nursing

staff with short-term contract nurses [156], and (iii) drawing nurses from less critical parts in the hospital during demand surges [156].

Methods: computer simulation [338, 365, 366, 367, 403, 519, 520, 590], heuristics [89, 156, 312], mathematical programming [32, 89, 156, 560], queueing theory [403], literature review [339, 428].

Facility layout. The facility layout concerns the positioning and organization of different physical areas in a facility. The aim is to determine the layout of the surgical care facility which maximizes the number of surgeries that can take place, given the budgetary and building constraints. A proper integration of the facility layout decision and the patient routing decision decreases costs and increases the number of patients operated [415]. For example, when patients are not anesthetized in the operating room, but in an adjacent induction room, patients can be operated with shorter switching times in between. In [428], contributions that model a facility layout decision for surgical care services are reviewed.

Methods: computer simulation [415], heuristics [468], literature review [428].

Tactical planning

Patient routing. A surgical process consists of multiple stages. We denote the composition and sequence of these stages as the route of a patient. The surgical process consists of a preoperative, perioperative and postoperative stage [262, 266, 482]. The preoperative stage involves waiting and anesthetic interventions, which can take place in induction rooms [415] or in the operating room [428]. The perioperative stage involves surgery in the operating room, and the postoperative stage involves recovery at a recovery ward [262]. Recovery can also take place in the operating room when a recovery bed is not immediately available [21]. Surgical patients requiring a bed are admitted to a (inpatient or outpatient) medical care unit before the start of the surgical process, where they return after the surgical process [333]. Efficient patient routes are designed with the objective to increase resource utilization [415].

Methods: computer simulation [415], heuristics [21], mathematical programming [21, 482], literature review [262, 428].

Capacity allocation. On the tactical level, resource capacities settled on the strategic level are subdivided over patient groups. The objectives of capacity allocation are to trade off patient access time and the utilization of surgical and postsurgical resources [58, 173, 262, 405, 560], to maximize contribution margin per hour of surgical time [99], to maximize the number of patients operated, and to minimize staff overtime [274]. Capacity allocation is a means to achieve an equitable distribution of access times [560]. Hospitals commonly allocate capacity through *block* scheduling [210, 262, 608]. Block scheduling involves the subdivision of operating time capacity in blocks that are assigned to patient groups [262, 266]. Capacity is allocated in three consecutive steps. First, patient groups are identified. Second, resource capacities, often in the form of operating time capacity, are subdivided over

the identified patient groups. Third, blocks of assigned capacity are scheduled to a specified date and time.

- *Patient group identification.* In general, patient groups are classified according to (sub)specialty, medical urgency, diagnosis or resource requirements. Identification by medical urgency distinguishes elective, urgent and emergent cases [99, 203, 262, 266]. Elective cases can be planned in advance, urgent cases require surgery urgently, but can incur a short waiting period, and emergency patients require surgery immediately [72, 99]. Examples of patient grouping by resource requirements are inpatients, day-surgery patients [266] and grouping patients by the equipment that is required for the surgery [156].

- *Time subdivision.* With the earlier mentioned objectives, operating time is subdivided over the identified patient groups based on expected surgery demand. This is often a politically charged and challenging task, since various surgical specialties compete for a profitable and scarce resource. What makes it even more complex is that hospital management and surgical specialties may have conflicting objectives [61]. When allocating operating time capacity to elective cases, a portion of total operating time capacity is reserved for emergency cases, which arrive randomly [233]. Staff overtime is the result when the reserved capacity is insufficient to serve all arriving emergency patients, but resource idle time increases when too much capacity is reserved, causing growth in elective waiting lists [72, 372, 373, 479, 656]. Capacity can be reserved by dedicating one or more operating rooms to emergency cases, or by reserving capacity in elective operating rooms [99, 371, 533].

- *Block scheduling.* In the last step of capacity allocation, a date and time are assigned to blocks of allocated capacity [40]. Several factors have to be considered in developing a block schedule. For example, (seasonal) variation in surgery demand, the number of available operating rooms, staff capacities, surgeon preferences, and material and equipment requirements [40, 513]. Block schedules are often developed to be cyclic, meaning the block schedule is repeated periodically. A (cyclic) block schedule is also termed a Master Surgical Schedule (MSS) [589]. Cyclic block schedules may not be suitable for rare elective procedures [262, 589]. For these procedures, capacity can be reserved in the cyclic block schedule [613], or non-cyclical plans may provide an outcome. When compared to cyclic plans, non-cyclic [156, 170, 171], or variable plans [339], increase flexibility, decrease staffing costs [156] and decrease patient access time [271, 339]. However, cyclic block schedules have the advantage that they make demand more predictable for surgical and downstream resources, such as the ICU and general wards, so that these resources can increase their utilization by anticipating demand more structurally [589].

In addition to block scheduling, the literature also discusses *open* scheduling and *modified block* scheduling. Open scheduling involves directly scheduling all patient groups in the total available operating time capacity, without subdividing this capacity first. Although open scheduling is more flexible than block scheduling, open

scheduling is rarely adopted in practice [61, 262], because it is not practical with regards to doctor schedules and increases competition for operating time capacity [405, 488]. Modified block scheduling is when only a fraction of operating time capacity is allocated by means of block scheduling [167, 262]. Remaining capacity is allocated and scheduled in a later stage, which increases flexibility to adapt the capacity allocation decision based on the latest information about fluctuating patient demand [262].

Capacity allocation decisions in surgical care services impact the performance of downstream inpatient care services [40, 42, 58, 99, 156, 405, 487, 591, 592, 593]. Variability in bed utilization and staff requirements can be decreased by incorporating information about the required inpatient beds for surgical cases in allocating surgical capacity [7, 40, 42, 255, 513, 588, 589]. In contributions that model downstream services, it is often the objective to level the bed occupancy in the wards or the ICU, to decrease the number of elective surgery cancelations [40, 99, 487, 513, 552, 560, 588, 589], or to minimize delays for inpatients waiting for surgery [652].

Methods: computer simulation [72, 167, 170, 171, 372, 479, 652], heuristics [40, 41, 42, 552, 606], Markov processes [233, 592, 593, 656], mathematical programming [40, 41, 42, 60, 61, 111, 156, 271, 372, 487, 513, 552, 560, 561, 588, 589, 652], queueing theory [656], literature review [58, 99, 262, 266, 339, 405, 468, 591, 608, 613].

Temporary capacity change. Available resource capacity could temporarily be adjusted in response to fluctuations in demand [403]. When additional operating time capacity is available, it can be allocated to a particular patient group, for example based on contribution margin [266, 613] or access times [560], or it can be proportionally subdivided between all patient groups [61, 560].

Methods: computer simulation [167], mathematical programming [61, 156, 560], literature review [266, 271, 613].

Unused capacity (re)allocation. Some time periods before the date of carrying out a settled block schedule, capacity allocation decisions may be reconsidered in order to reallocate capacity that remains unused [175, 266, 301] and to allocate capacity not allocated before (for example in *modified block* scheduling, discussed in *capacity allocation*). When unused capacity is released sufficiently early before the surgery time is planned, better quality reallocations are possible than when the unused capacity is released on the same day it is available [301]. Unused capacity is (re)allocated with the same objectives as the *capacity allocation* decision.

Methods: computer simulation [167, 175], heuristics [175], Markov processes [301], literature review [266].

Admission control. Admission control involves the rules according to which patients from different patient groups are selected to undergo surgery in the available operating time capacity. There is a strong reciprocal relation between admission control decisions and capacity allocation decisions: capacity allocation decisions demarcate the available operating time capacity for surgeries, and admission control decisions influence the required operating time capacity. Admission control has the

objective to balance patient service, resource utilization and staff satisfaction [58]. It is established by developing an admission plan that prescribes how many surgeries of each patient group to perform on each day, taking the block schedule into account [7]. Balancing the number of scheduled surgical cases throughout the week prevents high variance in utilization of involved surgical resources, such as operating rooms and recovery beds, and downstream inpatient care resources, such as ICU and general ward beds [6, 7, 40, 348, 405, 570]. Resource utilization can be improved by using call-in patients [58] and overbooking [72]. Call-in patients are given a time frame in which they can be called in for surgery when there is sufficient space available in the surgical schedule. Overbooking of patients involves planning more surgical cases than available operating time capacity to anticipate for no-shows [40]. Most patients requiring surgical care enter the hospital through the ambulatory care services. Although this makes admission control and capacity allocation policies for both ambulatory and surgical care services interdependent, not much literature is available on the interaction between ambulatory and surgical care services [591].

Methods: computer simulation [72, 165, 348, 570], Markov processes [443], mathematical programming [6, 7], literature review [58, 262].

Staff-shift scheduling. Shifts are hospital duties with a start and end time [91]. Shift scheduling deals with the problem of selecting what shifts are to be worked and how many employees should be assigned to each shift to meet patient demand [197]. The objective of shift scheduling is to generate shifts that minimize the number of staff hours required to cover the desired staffing levels [488]. The desired staffing levels are impacted by the capacity allocation decisions. Hence, integrated decision making for capacity allocation and staff-shift scheduling minimizes required staff [41]. Flexible shifts can improve performance [61, 166]. One example is staggered shift scheduling, which implies that employees have varying start and end times of shifts [468]. It can be used to plan varying, but adequate staffing levels during the day, and to decrease overtime [61, 166].

Methods: heuristics [162], mathematical programming [41, 88, 176], literature review [271, 488].

Offline operational planning

Staff-to-shift assignment. In staff-to-shift assignment, a date and time are given to a staff member to perform a particular shift. The literature on shift scheduling and assignment in healthcare mainly concerns inpatient care services [197], which we address in Section 2.6.

Methods: no papers found.

Surgical case scheduling. Surgical case scheduling is concerned with assigning a date and time to a specific surgical case. Availability of the patient, a surgeon, an anesthetist, nursing and support staff, and an operating room is a precondition [58]. Surgical case scheduling is an offline operational planning decision, since it results in an assignment of individual patients to planned resources and not in blueprints for assigning surgical cases to particular slots. The objectives of surgical

case scheduling are numerous: to achieve a high utilization of surgical and post-surgical resources, to achieve high staff and patient satisfaction, and to achieve low patient deferrals, patient cancelations, patient waiting time, and staff overtime [99, 156, 209, 333, 412, 482, 507, 533, 622]. The execution of a surgical case schedule is affected by various uncertainties in the preoperative stage duration, surgical procedure duration, switching time, postsurgical recovery duration, emergency patient interruption, staff availability, and the starting time of a surgeon [262, 482]. These uncertainty factors should be taken into account in surgical case scheduling.

Although surgical case scheduling can be done integrally in one step [21, 169, 170, 210, 405, 482, 507, 544], it is often decomposed in several steps. In the latter case, first, the planned length of a surgical case is decided. Second, a date and an operating room are assigned to a surgical case on the waiting list (also termed the 'advance scheduling' [405]). Third, the sequence of surgical cases on a specific day is determined [263, 405] (also termed the 'allocation scheduling' [405]). Fourth, starting times for each surgical case are determined. Below, we explain these four steps in more detail.

- *Planned length of a surgical case.* The planned length of a surgical case is the reserved operating time capacity in the surgical schedule for the surgical case duration, switching time and slack time. Surgical case duration, which is often estimated for each patient individually [464], is impacted by factors as the involved surgeon's experience, and the acuteness, sex, and age of the patient [163, 464]. Switching time between surgical cases includes cleaning the operating room, performing anesthetic procedures, or changing the surgical team [174]. Slack capacity is reserved as a buffer to deal with longer actual surgery durations than expected in advance [274]. When too little time is reserved, staff overtime and patient waiting time occur, and when too much time is reserved, resources incur idle time [174, 464, 622].
- *Assigning dates and operating rooms to surgical cases.* Dates and operating rooms are assigned to the elective cases on the surgical waiting list, following the settled admission control decisions [32, 208, 209, 274, 333, 416, 501]. The available blocks of operating time capacity are filled with elective cases. When too few cases are planned, utilization decreases, leading to longer waiting lists. Conversely, when too many cases are planned, costs increase due to staff overtime [72, 501]. Assigning dates and operating rooms to surgical cases can be done by assigning an individual surgical case, or by jointly assigning multiple cases to various possible dates and times. The latter is more efficient as more assignment possibilities can be considered [170].
- *Sequencing of surgical cases.* When the set of surgical cases for a day or for a block is known, the sequence in which they are performed still has to be determined. Factors to consider in the sequencing decision are doctor preference [262], medical or safety reasons [96, 333], patient convenience [96, 97], and resource restrictions [98]. Various rules for sequencing surgical cases are known [32, 96, 97, 264, 482, 501, 533]. In general, the traditional First-Come First-Served (FCFS) rule is outperformed by a Longest-Processing-Time-

First (LPTF) rule [58, 365, 367, 468]. When the variation of surgical case duration is known, sequencing surgical cases in the order of increasing case duration variation (i.e., lowest-variance-first) may yield further improvements [157, 622].

- *Assigning starting times to surgical cases.* The planned start time of each surgical case is decided [264]. This provides a target time for planning the presurgical and postsurgical resources, and for planning the doctor schedules [622]. The actual start time of a surgical case is impacted by the planned and actual duration of all preceding surgical cases [32, 622] and the completion time of the preoperative stage [172].

Emergency cases may play a significant role during the execution of the surgical case schedule [262]. Hence, incorporating knowledge about emergency cases, for example predicted demand, in surgical case scheduling decreases staff overtime and patient waiting time [72, 233, 371, 372, 373]. Often, surgical case scheduling is done in isolation. However, efficiency gains may be achieved by also considering decisions in other care services [96, 99, 122, 333, 482]. For example, without coordination with the ICU, a scheduled case may be rejected on its day of surgery due to a full ICU [482]. The contributions [21, 96, 122, 210, 312, 411, 444, 482, 520] do incorporate other care services, such as the patient wards and ICUs.

Methods: computer simulation [18, 72, 122, 164, 167, 169, 170, 173, 203, 264, 365, 367, 371, 372, 520, 561, 622], heuristics [18, 21, 98, 157, 163, 208, 210, 263, 264, 312, 371, 373, 416, 501, 507, 544, 587], Markov processes [233, 266, 443, 464], mathematical programming [21, 32, 96, 97, 98, 111, 122, 155, 156, 157, 208, 209, 210, 263, 333, 371, 372, 373, 412, 416, 479, 482, 501, 507, 533, 560], queueing theory [622], literature review [58, 99, 267, 405, 428, 468, 540].

Online operational planning

Emergency case scheduling. Emergency cases requiring immediate surgery are assigned to reserved capacity or to capacity obtained by canceling or delaying elective procedures [589]. It is the objective to operate emergency cases as soon as possible, but also to minimize disturbance of the surgical case schedule [267]. When emergency cases cannot be operated immediately, prioritizing of emergency cases is required to accommodate medical priorities or to minimize average waiting time of emergency cases [168, 482].

Methods: mathematical programming [168, 482], literature review [267].

Surgical case rescheduling. When the schedule is carried out, unplanned events, such as emergency patients, extended surgery duration and equipment breakdown may disturb the surgical case schedule [6, 412]. Hence, the surgical case schedule often has to be reconsidered during the day to mitigate increasing staff overtime, patient waiting time and resource idle time. Rescheduling may involve moving scheduled surgeries from one operating room to another and delaying, canceling or rescheduling surgeries [412].

Methods: mathematical programming [6, 412], literature review [266, 267].

Staff rescheduling. At the start of a shift, the staff schedule is reconsidered. Before and during the shift, the staff capacities may be adjusted to unpredicted demand fluctuations and staff absenteeism by using part-time, on-call nurses, staff overtime, and voluntary absenteeism [261, 483].

Methods: no papers found.

2.6 Inpatient care services

Inpatient care refers to care for a patient who is formally admitted (or 'hospitalized') for treatment and/or care and stays for a minimum of one night in the hospital [465]. Due to progress in medicine inpatient stays have been shortened, with many admissions replaced by more cost-effective outpatient procedures [466, 468]. Resource capacity planning has received much attention in the OR/MS literature, with capacity dimensioning being the most prominently studied decision.

Strategic planning

Regional coverage. At a regional planning level, the number, types and locations of inpatient care facilities have to be decided. To meet inpatient service demand, the available budget needs to be spent such that the population of each geographical area has access to a sufficient supply of inpatient facilities of appropriate nature and within acceptable distance [77]. Coordinated regional coverage planning between various geographical areas supports the realization of equity of access to care [56, 511]. To achieve this, local and regional bed occupancies need to be balanced with the local and regional probability of admission refusals resulting from a full census. The potential pitfall of deterministic approaches as used in [511] is that resource requirements are underestimated and thus false assurances are provided about the expected service level to patients [282].

Methods: computer simulation [282], mathematical programming [77, 511], queueing theory [56].

Service mix. The service mix is the set of services that healthcare facilities offer. Healthcare facilities that offer inpatient care services can provide a more complex mix of services and can accommodate patient groups with more complex diagnoses [540]. In general, the inpatient care service mix decision is not made at an inpatient care service level, but at the regional or hospital level, as it integrally impacts the ambulatory care facilities, the operating theater and the wards. This may be the reason that we have not found any references focusing on service mix decisions for inpatient care services in specific.

Methods: no papers found.

Case mix. Given the service mix decision, the types and volumes of patients that the facility serves need to be decided. The settled service mix decision restricts the decisions to serve particular patient groups. Patient groups can be classified based

on disease type, demographic information, and resource requirements [260]. In addition, whether patient admissions are elective or not is an influential characteristic on the variability of the operations of inpatient care services [578]. The case mix decision influences almost all other decisions, in particular the care unit partitioning and capacity dimensioning decisions [36].

Methods: computer simulation [260], heuristics [36, 578].

Care unit partitioning. Given the service and case mix decisions, the hospital management has to decide on the medical care units in which the inpatient care facility is divided. We denote this decision as care unit partitioning. It addresses both the question which units to create and the question which patient groups to consolidate in such care units. Each care unit has its designated staff, equipment and beds (in one or more wards). The objective is to guarantee care from appropriately skilled nurses and required equipment to patients with specific diagnoses, while making efficient use of scarce resources [36, 185, 186, 242, 282, 302, 527, 605].

First, the desirability of opening shared higher-level care units like Intensive Care Units (ICU) or Medium Care Units (MCU) should be considered [576]. Second, the general wards need to be specified. Although care unit partitioning is traditionally done by establishing a care unit for each specialty, or sometimes even more diagnosis specific [540], specialty-based categorization is not necessarily optimal. Increasingly, the possibilities and implications of consolidating inpatient services for care related groups is investigated to gain from the economies-of-scale effect, so-called 'pooling' [641]. For example, many hospitals merge the cardiac and thoracic surgery unit [255], or allow gynecologic patients in an obstetric unit during periods of low occupancy [430]. In such cases, the overflow rules need to be specified on the tactical level. For geriatric departments, it has to be decided whether to separate or consolidate assessment, rehabilitation and long-stay care [432, 434]. Also, multi-specialty wards can be created for patients of similar length of stay, such as day-care, short-, week- and long-stay units [527, 605], or for acute patients [315, 578]. Concentrating emergency activities in one area (a Medical Assessment Unit; MAU) can improve efficiency and minimize disruption to other hospital services [461]. One should be cautious when pooling beds for patient groups with diverging service level [255] or nursing requirements [374]. A combined unit would require the highest service and nurse staffing level for all patient groups. As a result, acceptable utilization may be lower than with separate units. Also, pooling gains should be weighed against possible extra costs for installing extra equipment on each bed [374]. To conclude, the question whether to consolidate or separate clinical services from a logistical point of view is one that should be answered for each specific situation, considering demand characteristics but also performance preferences and requirements [282]. Obviously, the care unit partitioning decision is highly interrelated with the *capacity dimensioning* decisions.

Methods: computer simulation [185, 186, 242, 282, 315, 527], heuristics [36, 374, 578], mathematical programming [461], queueing theory [255, 302, 430, 432, 434, 576, 641].

Capacity dimensioning. In conjunction with the care unit partitioning, the size of each care unit needs to be determined. Care unit size is generally expressed in the number of staffed beds, as this number is often taken as a guideline for dimensioning decisions for other resources such as equipment and staff.

- *Beds.* The common objective is to dimension the number of beds of a single medical care unit such that occupancy of beds is maximized while a predefined performance norm is satisfied [245, 457, 459, 500, 600, 634]. The typical performance measure is the percentage of patients that have to be rejected for admission due to lack of bed capacity: the admission refusal rate. Several other consequences of congested wards can be identified, all being a threat to the provided quality of care. First, patients might have to be transferred to another hospital in case of an emergency [128, 349, 424, 643]. Second, patients may (temporarily) be placed in less appropriate units, so-called misplacements [136, 185, 186, 255, 281, 286, 643]. Third, backlogs may be created in emergency rooms or surgical recovery units [124, 242, 255, 460, 461]. Fourth, elective admissions or surgeries may have to be postponed, by which surgical waiting lists may increase [13, 136, 245, 642, 643], which negatively impacts the health condition of (possibly critical) patients [570, 583]. Finally, to accommodate a new admission in critical care units, one may predischarge a less critical patient to a general ward [180, 628].

  The number of occupied beds is a stochastic process, because of the randomness in the number of arrivals and lengths of stay [355]. Therefore, slack capacity is required and thus care units cannot operate under 100% utilization [149, 255]. Often, inpatient care facilities adopt simple deterministic spreadsheet calculations, leading to an underestimation of the required number of beds [124, 136, 149, 277, 282]. Hospitals commonly apply a fixed target occupance level (often 85%), by which the required number of beds is calculated. Such a policy may result in excessive delays or rejections [27, 255, 282, 355, 457]. The desirable occupancy level should be calculated as a complex function of the service mix, the number of beds and the length of stay distribution [281, 282]. This non-linear relationship between number of beds, mean occupancy level and the number of patients that have to be rejected for admission due to lack of bed capacity is often emphasized [13, 149, 281, 286, 355, 457, 458, 500]. In determining the appropriate average utilization, the effect of economies of scale due to the so-called portfolio effect plays a role: larger facilities can operate under a higher occupancy level than smaller ones in trying to achieve a given patient service level [255, 282, 283, 355], since randomness balances out. However, possible economies of scope due to more effective treatment or use of resources should not be neglected [255]. Units with a substantial fraction of scheduled patients can in general operate under a higher average utilization [255]. The effect of variability in lengths of stay on care unit size requirements is shown to be less pressing than often thought by hospital managers [255, 583]. Reducing the average length of stay shows far more potential. For care units that have a demand profile with a clear time-dependent pattern, these effects are preferably

explicitly taken into account in modeling and decision making, to capture the seasonal [286, 404], day-of-week [180, 226, 286, 314] and even hour-of-day effects [38, 85, 124, 281]. This especially holds for units with a high fraction of emergencies admissions [540].

Capacity decisions regarding the size of a specific care unit can affect the operations of other units. Therefore, the number of beds needs to be balanced among interdependent inpatient care units [13, 83, 124, 125, 260, 282, 302, 315, 394, 424, 540]. Models that consider only a single unit neglect the possibility of admitting patients in a less appropriate care unit and thus the interaction between patient flows and the interrelationship between care units. Next to estimating utilization and the probability of admission rejections or delays, models that do incorporate multiple care units, also focus on the percentage of time that patients are placed in a care unit of a lower level or less appropriate care unit, or in a higher level care unit [19, 128, 236, 255, 394, 527]. The first situation negatively impacts quality of care as it can lead to increased morbidity and mortality [570] and the second negatively impacts both quality of care, as it may block admission of another patient, and efficient resource use [255, 527]. Some multi-unit models explicitly take the patient's progress through multiple treatment or recovery stages into account and try to dimension the care units such that patients can in each stage be placed in the care units that are most suitable regarding their physical condition [124, 128, 149, 205, 230, 235, 279, 284, 285, 315, 424, 527, 576].

- *Equipment.* In [605] it is stated that pooling equipment among care units can be highly beneficial. However, no references have been found explicitly focusing on this planning decision. This might be explained by the fact that the care unit partitioning and size decisions are generally assumed to be translatable to equipment capacity requirements. Therefore, many of the references mentioned under these decisions are useful for the capacity dimensioning of equipment.

- *Staff.* The highest level of personnel planning is the long-term workforce capacity dimensioning decision. This decision concerns both the number of employees that have to be employed, often expressed in the number of full time equivalents, and the mix in terms of skill categories [280, 460]. For inpatient care services it mainly concerns nursing staff. To deliver high-quality care, the workforce capacity needs to be such that an appropriate level of staff can be provided in the different care units in the hospital [197, 236]. In addition, holiday periods, training, illness and further education need to be addressed [91].

  Workforce flexibility is indicated as a powerful concept in reducing the required size of workforce [91, 152, 236, 540]. To adequately respond to patient demand variability and seasonal influences, it pays off to have substitution possibilities of different employee types, to use overtime, and to use part-time employees and temporary agency employees [540]. Just as with pooling bed capacity, economies of scale can be gained when pooling nursing staff among multiple care units. Nurses cross-trained to work in more than one unit can be placed in a so called 'float nurse pool' [91, 236, 374, 540]. Note that flexible staff can be

significantly more expensive [261]. Also, [382] indicates that to maintain the desired staff capacity, it is necessary to determine the long-term human resource planning strategies with respect to recruiting, promotion and training. To conclude, integrating the staff capacity dimensioning decision with the care unit size decision yields a significant efficiency gain [236].

Methods: computer simulation [13, 27, 124, 128, 136, 185, 186, 242, 260, 261, 277, 280, 281, 282, 283, 315, 349, 355, 424, 457, 458, 459, 460, 500, 527, 570, 600, 628, 634, 642, 643], heuristics [374], Markov processes [13, 85, 205, 230, 235, 284, 285, 286, 404], mathematical programming [152, 236, 280, 382, 394, 460, 461], queueing theory [19, 38, 83, 124, 125, 149, 180, 226, 245, 255, 279, 302, 314, 349, 394, 500, 576, 583], literature review [91, 197, 483, 540].

Facility layout. The facility layout concerns the positioning and organization of different physical areas in a facility. To determine the inpatient care facility layout, it needs to be specified which care units should be next to each other [468] and which care units should be close to other services like the surgical, emergency and ambulatory care facilities [93]. Ideally, the optimal physical layout of an inpatient care facility is determined given the decisions on service mix, case mix, care unit partitioning and care unit size. However, in practice, it often happens vice versa: physical characteristics of a facility constraint service mix, care unit partitioning and care unit size decisions [93, 605]. Newly-built hospitals are preferably designed such that they support resource pooling and have modular spaces so that they are as flexible as possible with respect to care unit partitioning and dimensioning [605].

Methods: computer simulation [93], heuristics [468], mathematical programming [93].

Tactical planning

Bed reallocation. Given the strategic decision making, tactical resource allocation needs to ensure that the fixed capacities are employed such that inpatient care is provided to the right patient groups at the right time, while maximizing resource utilization. Bed reallocation is the first step in tactical inpatient care service planning. Medium-term demand forecasts may expose that the care unit partitioning and size decisions fixed at the strategic level are not optimal. If the ward layout is sufficiently flexible, a reallocation of beds to units or specialties based on more specific demand forecast can be beneficial [36, 281, 606]. In addition, demand forecasts can be exploited to realize continuous reallocation of beds in anticipation for seasonality in demand [342]. To this end, hospital bed capacity models should incorporate monthly, daily and hourly demand profiles and meaningful statistical distributions that capture the inherent variability in demand and length of stay [277]. When reallocating beds, the implications for personnel planning, and involved costs for changing bed capacity, should not be overlooked [12].

Methods: computer simulation [281, 342], heuristics [36, 606], mathematical programming [12], queueing theory [342].

Temporary bed capacity change. To prevent superfluous staffing of beds, beds can

temporarily be closed by reducing staff levels [255]. This may for instance be in response to predicted seasonal or weekend demand effects [277, 283]. The impact of such closings on the waiting lists at referring outpatient clinics and the operating room is studied by [641, 642]. Temporary bed closings may also be unavoidable as a result of staff shortages [424]. In such cases hospitals can act pro-actively, to prevent bed closings during peak demand periods [36].

Methods: computer simulation [277, 283, 424, 642], heuristics [36], queueing theory [255, 641].

Admission control. To provide timely access for each different patient group, admission control prescribes the rules according to which various patients with different access time requirements are admitted to nursing wards. At this level, patients are often categorized in elective, urgent and emergency patients. Admission control policies have the objective to match demand and supply such that access times, rejections, surgical care cancelations and misplacements are minimized while bed occupancy is maximized. The challenge is to cope with variability in patient arrivals and length of stay. Smoothing patient inflow, and thus workload at nursing wards, prevents large differences between peak and non-peak periods, and so realizes a more efficient use of resources [7, 277, 607].

Patient resource requirements are another source of variability in the process of admission control. Most references only focus on maximizing utilization of bed resources. This may lead to extreme variations in the utilization of other resources like diagnostic equipment and nursing staff [540]. Also, as with temporarily closing of beds, possible effects of admission control policies on the waiting lists at referring outpatient clinics and the operating room should not be neglected [531]. Admission control policies can be both static (following fixed rules) and dynamic (changing rules responding to the actual situation).

- *Static bed reservation.* To anticipate for the estimated inflow of other patient groups, two types of static bed reservation can be distinguished. The first is refusing admissions of a certain patient type when the bed census exceeds a threshold. For example, to prevent the rejection of emergent admission requests, an inpatient care unit may decide to suspend admissions of elective patients when the number of occupied beds reaches a threshold [198, 227, 327, 343, 424, 430, 500, 531]. As such, a certain number of beds is reserved for emergency patients. This reservation concept is also known as 'earmarking'. Conversely, [350, 570] indicate that earmarking beds for elective postoperative patients can minimize operating room cancelations. In the second static level the number of reserved beds varies, for example per weekday. Examples of such a policy are provided in [55, 577] where for each work day a maximum reservation level for elective patients is determined.
- *Dynamic bed reservation.* Dynamic bed reservation schemes take into account the actual 'state' of a ward, expressed in the bed census per patient type. Together with a prediction of demand, the reservation levels may be determined for a given planning horizon [356] or it may be decided to release reserved beds when demand is low. Examples of the latter are found in [350], where bed reser-

vations for elective surgery are released during weekend days, and [39], where admission quota are proposed per weekday. In [292], an extension to dynamic reservation is proposed which concerns calling in semi-urgent patients from an additional waiting list on which patients are placed who needs admission within 1–3 days.

- *Overflow rules.* In addition to the bed reservation rules, overflow rules prescribe what happens in the case that all reserved beds for a certain patient type are occupied. In such cases, specific overflow rules prescribe which patient types to place in which units [282]. Generally, patients are reassigned to the correct treatment area as soon as circumstances permit [540]. By allowing overflow and setting appropriate rules, the benefits of bed capacity pooling are utilized (see *capacity dimensioning: care unit size*), while the alignment of patients with their preferred bed types is maximized [424]. Various references focus on predicting the impact of specific overflow rules [248, 282, 302, 424, 527].
- *Influence surgical schedule.* For many inpatient care services the authority on admission control is limited due to the high dependency on the operating room schedule (see *surgical care services*). By adjusting the surgical schedule, extremely busy and slack periods can be avoided [7, 36, 180, 185, 203, 248, 255, 277, 560, 592, 593, 606, 607, 643] and cancelation of elective surgeries can be avoided [348]. In practice, the operating room planning is generally done under the assumption that a free bed is available for postoperative care [350], which may result in surgery cancelations. Therefore, both for inpatient and surgical care services coordinated planning is beneficial [6, 277].

Methods: computer simulation [6, 185, 203, 248, 277, 282, 348, 350, 424, 500, 527, 560, 570, 607, 643], heuristics [36, 606], Markov processes [55, 198, 292, 302, 356, 592, 593], mathematical programming [6, 7, 39, 560], queueing theory [39, 180, 227, 255, 327, 343, 430, 531, 577]

Staff-shift scheduling Shifts are hospital duties with a start and end time [91]. Shift scheduling deals with the problem of selecting what shifts are to be worked and how many employees should be assigned to each shift to meet patient demand [197, 346]. For inpatient care services, it generally concerns the specification of 24-hours-a-day-staffing levels divided in a day, evening and night shift, during which demand varies considerably [91, 197]. Typically, this is done for a period of one or two months [483]. Staffing levels need to be set both for each care unit's dedicated nurses and for flexible staff in floating pools [374]. Also, [152, 261] investigate the potential of on-call nurses who are planned to be available during certain shifts and only work when required.

The first step in staff shift scheduling is to determine staffing requirements with a demand model [197, 271, 346, 553], based on which the bed occupancy levels [540] and medical needs are forecasted [374]. The second step is to translate the forecasted demand in workable shifts and in the number of nurses to plan per shift, taking into account the staff resources made available at the strategic decision level [618]. Often, nurse-to-patient ratios are applied in this step [261], which are

assumed to imply acceptable levels of patient care and nurse workload [644]. To improve the alignment of care demand and supply, shift scheduling is preferably coordinated with scheduled admissions and surgeries [483], which also helps avoiding high variation in nurse workload pressure [41].

Methods: computer simulation [261], heuristics [374], mathematical programming [41, 152, 618, 644], queueing theory [553], literature review [91, 197, 271, 346, 483, 540].

Offline operational planning

Admission scheduling. Governing the rules set by tactical admission control policies, on the operational decision level the admission scheduling determines for a specific elective patient the time and date of admission. We found one reference on this decision: [134] presents a scheduling approach to schedule admissions for a short-stay inpatient facility that only operates during working days, which takes into account various resource availabilities such as beds and diagnostic resources. We suggest two reasons for the lack of contributions on this decision. First, when admission control policies are thoroughly formulated, admission scheduling is fairly straightforward. Second, as described before, for postoperative inpatient care services authority of admission planning is generally at the surgical care services [606].

Methods: mathematical programming [134].

Patient-to-bed assignment. Together with the admission scheduling decision, an elective patient needs to be assigned to a specific bed in a specific ward. Typically, this assignment is carried out a few days before the effective admission of the patient. The objective is to match the patient with a bed, such that both personal preferences (for example a single or twin room) and medical needs are satisfied [110, 153]. An additional objective may be to balance bed occupancy over different wards.

Methods: heuristics [110, 153], mathematical programming [110, 153].

Discharge planning. Discharge planning is the development of an individualized discharge plan for a patient prior to leaving the hospital. It should ensure that patients are discharged from the hospital at an appropriate time in their care and that, with adequate notice, the provision of other care services is timely organized. The aim of discharge planning is to reduce hospital length of stay and unplanned readmission, and improve the coordination of services following discharge from the hospital [529]. As such, discharge planning is highly dependent on availability downstream care services, such as rehabilitation, residential or home care. Therefore, a need is identified for integrated coherent planning across services of different healthcare organizations [598, 623]. Patients whose medical treatment is complete but cannot leave the hospital are often referred to as 'alternative level of care patients' or 'bed blockers' [591, 623]. Also in discharge planning it is worthwhile to anticipate for seasonality effects.

Methods: computer simulation [598], queueing theory [623], literature review [529].

Staff-to-shift assignment. Staff-to-shift assignment deals with the allocation of staff members to shifts over a period of several weeks [197]. The term 'nurse rostering' is also often used for this step in inpatient care services personnel planning [91, 118]. The objective is to meet the required shift staffing levels set on the tactical level, while satisfying a complex set of restrictions involving work regulations and employee preferences [52, 91, 118, 332, 346, 579]. Night and weekend shifts, days off and leaves have to be distributed fairly [483, 540, 644] and as much as possible according to individual preferences [52, 197]. In most cases, to compose a roster for each individual, first sensible combinations or patterns of shifts are generated (cyclic or non-cyclic), called 'lines-of-work', after which individuals are assigned to these lines-of-work [197]. Sometimes, staff-to-shift assignment is integrated with staff-shift scheduling [91, 644]. 'Self-scheduling' is an increasingly popular concept aimed at increased staff satisfaction which allows staff members to first propose individual schedules, which are taken as starting point to create a workable schedule that satisfies the staffing level requirement set on the tactical level [508].

Methods: heuristics [52, 579], mathematical programming [52, 332, 508, 579, 644], literature review [91, 118, 197, 346, 483, 540].

Online operational planning

Elective admission rescheduling. Based on the current status of both the patient and the inpatient care facility, it has to be decided whether a scheduled admission can proceed as planned. Circumstances may require postponing or canceling the admission, to reschedule it to another care unit, or to change the bed assignment. Various factors will be taken into consideration such as severity of illness, age, expected length of stay, the probable treatment outcome, the (estimated) bed availability, and the conditions of other patients (in view of the possibility of predischarging an other patient) [349, 398, 530]. This decision is generally made on the planned day of admission or a few days in advance. Rescheduling admissions can have a major impact on the operations at the surgical theater [349].

Methods: computer simulation [349], heuristics [398], queueing theory [349, 530].

Acute admission handling. For an acute admission request it has to be decided whether to admit the emergency patient and if so to which care unit, which bed, and on what notice. The tactical admission control rules act as guideline. As with rescheduling elective admissions, the status of both the patient and the inpatient care facility are taken into account [349, 530]. In [349], it is calculated how long the waiting will be if the patient is placed on 'the admission list' and [530] proposes and evaluates an admission policy to maximize the expected incremental number of lives saved from selecting the best patients for admission to an ICU.

Methods: computer simulation [349], queueing theory [349, 530].

Staff rescheduling. At the start of a shift, the staff schedule is reconsidered. Based on the severity of need in each care unit, the float nurses and other flexible employees are assigned to a specific unit and a reassignment of dedicated nurses may

also take place [91, 540]. In addition, before and during the shift, the staff capacities among units may be adjusted to unpredicted demand fluctuations and staff absenteeism by using float, part-time, on-call nurses overtime, and voluntary absenteeism [261, 483].

Methods: computer simulation [261], mathematical programming [489], literature review [91, 483, 540].

Nurse-to-patient assignment. At the beginning of each shift, each nurse is assigned to a group of patients to take care for. This assignment is done with the objective to provide each patient with an appropriate level of care and to balance workloads [447, 546]. Distributing work fairly among nurses improves the quality of care [447]. Generally, the assignment has to satisfy prespecified nurse-to-patient ratios [489]. Additionally, when patient conditions within one care unit can differ considerably, for each specific patient an estimate of the severity of the condition (and thereby expected workload) is made, in most cases on the basis of a certain severity scoring system [447]. In [489], it is explicitly taken into account that patient conditions, and therefore care needs, can vary during a shift. They state that it is preferred to also decide at the beginning of each shift to which nurse(s) unanticipated patients will be assigned.

Methods: computer simulation [546], heuristics [447], mathematical programming [447, 489].

Transfer scheduling. Throughout the inpatients' stay, the transfer scheduling is done to the appropriate inpatient care unit or to other areas within the hospital for treatments or diagnoses [483]. Transfer scheduling includes the planning of transportation. Transfer scheduling is often postponed until the time an already occupied bed is requested by a new patient. However, in [563] it is concluded that when relocation of patients is done proactively, admission delays for other patients can significantly be reduced, which has a positive effect on both quality and efficiency.

Methods: Markov processes [563].

2.7 Discussion

This chapter has introduced a taxonomy to identify, break down and classify decisions to be made in the managerial field of healthcare resource capacity planning and control. It has provided a structured overview of the planning decisions in six identified care services and the corresponding state of the art in OR/MS literature. Having done this, we aim for an impact that is threefold. First, we aim to support healthcare professionals in improved decision making. Second, we aim to inspire OR/MS researchers in formulating future research objectives and to position their research in a hierarchical framework. Third, we aim for interconnecting healthcare professionals and OR/MS researchers so that the potential of OR/MS can be discovered and exploited in improving healthcare delivery performance.

The vertical axis in our taxonomy represents the hierarchical nature of decision making in healthcare organizations. Aggregate decisions are made in an early

stage, and finer granularity is added in later stages when more detailed information becomes available. The observed literature explicitly substantiates the relations between planning decisions both within and between hierarchical levels. Planning decisions on higher levels shape decision making on lower levels by imposing restrictions. Performance on lower levels concerns feedback about the realization of higher level objectives, thereby potentially impacting decision making on higher levels. We have seen many examples of these interactions in our review. Incorporating flexibility in planning reduces restrictions imposed by decisions settled in higher levels on lower level decision making. Increased planning flexibility involves specifying and adjusting planning decisions closer to the time of actual healthcare delivery, thereby giving the opportunity to incorporate more detailed and accurate information in decision making. The observed contributions that incorporate planning flexibility provide opportunities to improve the response to fluctuations in patient demand and thus to improve performance.

Although organized by different organizations, the healthcare delivery process from the patient's perspective generally is a composition of several care services. A patient's pathway typically includes several care stages performed by various healthcare services. The effectiveness and efficiency of healthcare delivery is a result of planning and control decisions made for the care services involved in each care stage. The quality of decision making in each care service depends on the information available from and the restrictions imposed by other care services. Therefore, in the perspective of the presented taxonomy, in addition to the vertical interaction, a strong horizontal interaction can be recognized. Suboptimization is a threat when these decisions are taken in isolation. At various points in our overview, we have observed that an integrated decision making approach is beneficial. Such an integration is not straightforward as it also emerged that different care services may have conflicting objectives. Our categorization of planning decisions in Sections 2.4-2.6 based on the taxonomy presented in Section 2.2, enables identification of interactions between different care services, allows detection of conflicting objectives, and helps to discover opportunities for coordinated decision making.

Due to the segmented organizational structure of healthcare delivery, also the OR/MS literature has initially focused predominantly on autonomous, isolated decision making. Such models ignore the inherent complex interactions between decisions for different services in different organizations and departments. In 1999, the survey [339] identified a void in OR/MS literature focusing on integrated healthcare systems. The level of complexity of such models is depicted as main barrier. In 2010, the survey [591], reviewing OR/MS models that encompass patient flows across multiple departments, addressed the question whether this void has since been filled. The authors conclude that the lack of models for complete healthcare processes still existed. Although a body of literature focusing on two-departmental interactions was identified, very few contributions were found on complete hospital interactions, let alone on complete healthcare system interactions. The present review of the literature reconfirms these observations.

To conclude, the specification of planning decisions in our taxonomy allows for

identifying relations within and between hierarchical levels. Recognizing and incorporating these relationships in decision making improves healthcare delivery performance. Creating more planning flexibility in decision making demonstrates great potential. By specifying and adjusting planning decisions closer to the time of actual healthcare delivery, more detailed and accurate information can be incorporated, providing opportunities to adjust planning decisions to better match care supply and demand. Furthermore, integrated decision making for multiple care services along a care chain shows great potential. With the growing awareness of the potential benefit of such integrated decision making, an increase in the number of publications in which integrated models are presented is noticeable [99, 591]. However, it remains a challenge for OR/MS researchers to develop integral models that on the one hand provide an extensive healthcare system scope, while on the other hand incorporating a satisfactory level of detail and insight. Summarizing, for the sake of patient centeredness and cost reductions required by societal voices and pressures, we claim that both healthcare professionals and OR/MS researchers should address coordinated and integrated decision making for interrelated planning decisions, should explore the opportunities of increased flexibility, and should take an integral care chain perspective.

2.8 Appendix

2.8.1 Search terms

This appendix presents the search terms that were used to identify the literature base set in literature search method described in Section 2.3.

Care service	*Search terms*
Ambulatory care	“outpatient clinic$” OR “outpatient facilit*” OR “outpatient care” OR “ambulatory care” OR “ambulatory health center$” OR “diagnostic service$” OR “diagnostic facilit*” OR “radiology”
Emergency care	“emergenc*” OR “acute” OR “accident” OR “ambulance” AND “health”
Surgical care	“operating room$” OR “operating theat*” OR “surgery scheduling” OR “operating suite” OR “surgical” OR “surger*”
Inpatient care	“bed$” OR “intensive care” OR “ward$” AND “hospital”
Home care	“home care” OR “home health care” OR “home-care” OR “home-health-care” OR “home-health care” OR “home healthcare”
Residential care	“nursing home$” OR “mental care” OR “rehabilitation cent*” OR “rehabilitation care” OR “long-term care" OR (“retirement” OR “geriatric” OR “elderly” AND "health")

2.8.2 Overview tables of the identified planning decisions

This appendix displays the overview tables of the identified planning decisions and the applied OR/MS methods for each of the six care services. In the overview tables, the following acronyms are used when referring to the methods:

Method	*Acronym*
Computer simulation	CS
Heuristics	HE
Markov processes	MV
Mathematical programming	MP
Queueing theory	QT
Literature review	LR

Ambulatory care services

Level	Planning decision	CS	HE	MV	MP	QT	LR
Strategic	Regional coverage	[425, 505, 543, 562]	[2, 181]				[540]
	Service mix						
	Case mix	[548]			[539]		
	Panel size	[543]				[256]	
	Capacity dimensioning:						
	– consultation rooms	[321, 547, 548]				[321]	[339, 540]
	– staff	[425, 506, 543, 547, 548, 626]		[619]	[539]	[43]	[339, 540]
	– consultation time capacity	[189, 193]				[139, 193]	[339, 540]
	– equipment	[225, 425, 562]					[339, 540]
	– waiting room	[548]					[339, 540]
	Facility layout		[468]				
Tactical	Patient routing	[113, 225, 321, 425, 543]				[321, 655]	
	Capacity allocation	[604]		[265, 539]			[608]
	Temporary capacity change	[193, 604]					
	Access policy	[20, 212, 472, 502, 599]		[472]		[504, 655]	
	Admission control	[604]	[237]	[232, 237]	[337, 490, 492]		
	Appointment scheduling	[22, 100, 105, 160, 189, 212, 213, 278, 308, 309, 345, 368, 390, 400, 425, 463, 502, 548, 608, 609, 624, 629]	[100, 340]	[220, 257, 340, 357, 396, 451, 542]	[100, 155, 503]	[73, 139, 179, 363, 386, 503, 608, 655]	[104, 267, 339, 540]
	Staff-shift scheduling	[478]			[88]		[91, 197, 271, 468]
Offline operational	Patient-to-appointment assignment:						
	– single appointment		[120, 621]	[268, 474, 621]			
	– combination appointments		[481]				
	– appointment series				[132, 133, 135]		
	Staff-to-shift assignment				[337]		[271]
Online operational	Dynamic patient (re)assignment	[497]		[146, 257, 396]	[146]		
	Staff rescheduling						

Emergency care services

Level	Planning decision	CS	HE	MV	MP	QT	LR
Strategic	Regional coverage						
	– emergency care centers	[77]	[44]	[29, 313]	[94, 244, 313, 336, 499, 566]	[44]	[252, 336, 395, 483, 499]
	– ambulances	[77, 196, 214, 224, 240, 276, 323, 498, 514, 549, 645]	[30, 44, 46, 195, 234, 322]		[26, 44, 46, 47, 48, 53, 145, 187, 196, 224, 241, 276, 322, 336, 498, 499, 535, 549]	[44, 234, 322, 380, 413, 535]	[84, 252, 336, 395, 483, 499]
	Service mix						
	Ambulance districting	[240, 514]	[44]		[44]	[44, 101, 380]	
	Capacity dimensioning:						
	– ambulances	[48, 196, 224, 323, 498, 514, 645]	[46]		[46, 47, 196, 498]	[535, 559]	
	– waiting room					[126]	[475]
	– treatment rooms	[108, 376]				[126]	[475]
	– emergency wards	[23, 381, 460]			[460, 461]	[126]	
	– equipment	[108]				[126]	[475]
	– staff	[75, 215, 376, 460, 651]			[460, 461]	[259]	[77, 339, 475]
	Facility layout	[651]	[468]				[475]
Tactical	Patient routing	[75, 108, 215, 375, 426, 596]				[126, 429]	[339, 475]
	Admission control	[75, 108]				[429]	
	Staff-shift scheduling	[323, 536, 537, 651]	[536, 537]			[253, 254, 259]	[271, 339, 475]
Offline operational	Staff-to-shift assignment		[103]		[28, 35, 103, 151, 195]		
Online operational	Ambulance dispatching	[16, 388, 397, 645]	[388]		[397]	[559]	
	Facility selection	[514]					
	Ambulance routing						
	Ambulance relocation	[16, 231, 645]		[427, 518]	[231]		[84]
	Treatment planning and prioritization	[108, 215]					
	Staff rescheduling	[651]			[460]		

Surgical care services

Level	Planning decision	CS	HE	MV	MP	QT	LR
Strategic	Regional coverage	[71]			[513]		
	Service mix						
	Case mix	[338]			[59, 317]		[262]
	Capacity dimensioning:						
	– operating rooms	[520]			[32]		[339]
	– operating time capacity	[338, 403, 519, 590]			[560]	[403]	[428]
	– presurgical rooms						
	– recovery wards	[365, 366, 367, 519, 520]					[339]
	– ambulatory surgical ward						
	– equipment						
	– staff		[89, 156, 312]		[89, 156]		[339]
	Facility layout	[415]	[468]				[428]
Tactical	Patient routing	[415]	[21]		[21, 482]		[262, 428]
	Capacity allocation	[72, 167, 170, 171, 372, 479, 652]	[40, 41, 42, 552, 606]	[233, 592, 593, 656]	[40, 41, 42, 60, 61, 111, 156, 271, 372, 487, 513, 552, 560, 561, 588, 589, 652]	[656]	[58, 99, 262, 266, 339, 405, 468, 591, 608, 613]
	Temporary capacity change	[167]			[61, 156, 560]		[266, 271, 613]
	Unused capacity (re)allocation	[167, 175]	[175]	[301]			[266]
	Admission control	[72, 165, 348, 570]		[443]	[6, 7]		[58, 262]
	Staff-shift scheduling		[162]		[41, 88, 176]		[271, 488]
Offline operational	Staff-to-shift assignment						
	Surgical case scheduling	[18, 72, 122, 164, 167, 169, 170, 173, 203, 264, 365, 367, 371, 372, 520, 561, 622]	[18, 21, 98, 157, 163, 208, 210, 263, 264, 312, 371, 373, 416, 501, 507, 544, 587]	[233, 266, 443, 464]	[21, 32, 96, 97, 98, 111, 122, 155, 156, 157, 208, 209, 210, 263, 333, 371, 372, 373, 412, 416, 479, 482, 501, 507, 533, 560]	[622]	[58, 99, 267, 405, 428, 468, 540]
Online operational	Emergency case scheduling				[168, 482]		[267]
	Surgical case rescheduling				[6, 412]		[266, 267]
	Staff rescheduling						

Inpatient care services

Level	Planning decision	CS	HE	MV	MP	QT	LR
Strategic	Regional coverage	[282]			[77, 511]	[56]	
	Service mix						
	Case mix	[260]	[36, 578]				
	Care unit partitioning	[185, 186, 242, 282, 315, 527]	[36, 374, 578]		[461]	[255, 302, 430, 432, 434, 576, 641]	
	Capacity dimensioning:						
	– beds	[13, 27, 124, 128, 136, 185, 186, 242, 260, 277, 281, 282, 283, 315, 349, 355, 424, 460, 457, 458, 459, 500, 527, 570, 600, 628, 634, 642, 643]		[13, 85, 205, 230, 235, 284, 285, 286, 404],	[236, 394, 460, 461]	[19, 38, 83, 124, 125, 149, 180, 226, 245, 255, 279, 302, 314, 349, 394, 500, 576, 583]	
	– equipment						
	– staff	[261, 280, 460]	[374]	[152, 236, 280, 382, 460]			[91, 197, 483, 540]
	Facility layout	[93]	[468]		[93]		
Tactical	Bed reallocation	[281, 342]	[36, 606]		[12]	[342]	
	Temporary bed capacity change	[277, 283, 424, 642]	[36]			[255, 641]	
	Admission control:						
	– static bed reservation	[350, 424, 500, 570]		[55, 198]		[227, 327, 343, 430, 531, 577]	
	– dynamic bed reservation	[350]		[292, 356]	[39]	[39]	
	– overflow rules	[248, 282, 424, 527]		[302]			
	– influence surgical schedule	[6, 185, 203, 248, 277, 348, 350, 560, 607, 643]	[36, 606]	[592, 593]	[6, 7, 560]	[180, 255]	
	Staff-shift scheduling	[261]	[374]		[41, 152, 618, 644]	[553]	[91, 197, 271, 346, 483, 540]
Offline operational	Admission scheduling				[134]		
	Patient-to-bed assignment		[110, 153]	[110, 153]			
	Discharge planning	[598]				[623]	[529]
	Staff-to-shift assignment		[52, 579]		[52, 332, 508, 579, 644]		[91, 118, 197, 346, 483, 540]

Level	Planning decision	CS	HE	MV	MP	QT	LR
Online operational	Elective adm. rescheduling	[349]	[398]			[349, 530]	
	Acute admission handling	[349]				[349, 530]	
	Staff rescheduling	[261]			[489]		[91, 483, 540]
	Nurse-to-patient assignment	[546]	[447]		[447, 489]		
	Transfer scheduling			[563]			

Home care services

Level	Planning decision	CS	HE	MV	MP	QT	LR
Strategic	Placement policy		[648]	[379]	[112]		[45]
	Regional coverage						[45]
	Service mix						[45]
	Case mix						[45]
	Panel size				[147]		
	Districting		[57]				[45]
	Capacity dimensioning:						
	– staff	[585]		[275]		[92]	[45]
	– equipment	[470]					[45]
	– fleet vehicles						[45]
Tactical	Capacity allocation:						
	– patient group identification					[92]	[45]
	– time subdivision		[62, 370]		[147]		
	Admission control			[379]	[147, 303]	[92]	
	Staff-shift scheduling		[370]				[45]
Offline operational	Assessment and intake		[538, 648]	[379]	[112, 202, 303]		[45]
	Staff-to-shift assignment		[49, 202]		[49, 202]		[45]
	Visit scheduling:						
	– short-term care plan		[37, 201, 202]	[379]	[201, 202, 303]		[45]
	– staff-to-visit assignment		[37, 49, 201, 202]	[379]	[49, 82, 201, 202, 303]		[45]
	– route creation		[37, 49, 80, 81, 201, 202]		[49, 80, 81, 82, 112, 201, 202]		[45]
Online operational	Visit rescheduling		[37, 201, 202, 567]		[37, 201, 202, 567]		

Residential care services

Level	Planning decision	CS	HE	MV	MP	QT	LR
Strategic	Placement policy	[431, 585]	[33, 648]	[123, 206, 207, 229, 275, 326, 420, 421, 422, 423, 431, 432, 433, 434, 435, 442, 477, 526, 528, 555, 556, 557, 558, 612, 646, 647]		[249]	
	Regional coverage				[77, 137, 184]		[483]
	Case mix		[221]				
	Capacity dimensioning:						
	– beds	[158, 190, 354, 431, 471, 585]		[123, 206, 207, 229, 235, 275, 326, 420, 421, 422, 423, 431, 432, 433, 434, 442, 471, 477, 526, 528, 555, 556, 557, 558, 612, 646, 647]		[117, 246, 249, 354, 623]	
	– staff	[158]					
Tactical	Admission control	[471]			[230, 389, 471]		
Offline operational	Treatment scheduling				[517]		

The OR/MS literature directed to residential care services showed a low variety in addressed planning decisions. The dynamics of residential care services, although on a slower time scale, are similar to that of inpatient care services. Therefore, most planning decisions and insights described under inpatient care services also apply to residential care services. These are the reasons that we have chosen for residential care services, as opposed to the other care services, to only present planning decisions for which we found references.

Part III

Facilitating the One-Stop Shop Principle

CHAPTER 3

Balancing Appointments and Walk-ins

3.1 Introduction

Developing appointment schedules for outpatient care facilities that process both patients with and without an appointment is challenging, as it requires planning and scheduling on different time scales. A well-designed appointment system comprises an efficient day appointment schedule and provides timely access. This chapter is motivated by challenges faced by hospital outpatient clinics that serve patients on a walk-in basis. Most of these clinics also have a limited number of appointment slots. There are various organizational (e.g., fixed slots for patients in a care pathway, patients with long travel time to the hospital, children) and medical (e.g., local anesthesia or contrast fluid required) reasons to give a patient an appointment. We introduce a method to design appointment schedules for such facilities.

Advantages of a walk-in system are a higher level of accessibility and more freedom for patients to choose the date and time of their hospital visit. Disadvantages are a possible highly variable demand and as a consequence low utilization and high waiting time (the time between the physical arrival at the facility and the start of consultation and/or treatment). The advantage of an appointment system is that workload can be dispersed, while it has the disadvantage of a potentially long access time (the time between the day of the appointment request and the appointment date). Since prolonged access times result in a delay of treatment, deterioration of health condition is a serious risk [449]. Allowing patients to walk in effectively reduces access times to zero, and thus increases quality of care. In addition, healthcare facilities typically aim to guarantee a certain service level with respect to the access time for patients with an appointment.

The challenge in a mixed system is thus to balance access time for appointment patients and waiting time for walk-in patients. To achieve this, we develop a methodology that schedules appointments when the expected walk-in demand is low. To smoothen the system, in periods of high demand part of the walk-in patients is offered an appointment at a later moment. Of course, this is undesirable since it increases access time and may involve an additional clinic visit. Walk-in demand [20, 126] and demand for appointments requests [633] are often cyclic; therefore, we develop a cyclic appointment schedule. Appointment scheduling has received considerable attention in the literature (see Chapter 2), as opposed to the

development of models that relate access and waiting time [267].

Our contribution is a methodology that incorporates scheduled and unscheduled arrivals and maximizes the number of unscheduled patients served on the day of arrival, while satisfying a pre-specified access time norm for scheduled patients. We model the unscheduled arrivals with a stochastic non-stationary arrival process and incorporate balking behavior. The scheduled patients have priority, may not show up, and appointment requests are assumed to arrive according to a cyclic pattern. To account for the cyclic arrivals, the appointment schemes we develop are also cyclic, where the cycle is a repeating sequence of days. The cycle length can, for instance, be a week or a month. The cyclic appointment schedule (CAS) specifies a capacity cycle (the maximum number of patients that can be scheduled on each day of the cycle) and a day schedule (the maximum number of patients to be scheduled per time slot on each day). Access and waiting time are measured on different time scales, since access time is counted between days and waiting time during a day.

To facilitate the two time scales, our approach consists of decomposing the appointment planning process and the service process during the day. For both processes we propose an analytical evaluation model. The first model determines the access time for scheduled patients for any given capacity cycle. The second model determines the expected number of unscheduled patients that cannot be seen on the day of arrival. The two models are linked by an iterative algorithm that stops when the CAS is found in which the fraction of unscheduled patients seen on the day of arrival is maximized, given that the restriction on the access time is satisfied. A numerical example of a small problem instance demonstrates the potential of the methodology. In this example complete enumeration is applied to find optimal day schedules. Our future research will aim at incorporating heuristics to quickly find (close to) optimal day schedules, so that larger problem sizes can be tackled. Finding an optimal day schedule is not straightforward and a field of research on its own [104, 267].

This chapter is organized as follows. Section 3.2 provides a literature review. In Section 3.3, we give an introduction to the methodology and provide a formal problem description. Sections 3.4, 3.5, and 3.6 respectively present the access and day process evaluation models, and the algorithm. Section 3.7 describes the numerical example, followed by the discussion and conclusions in Section 3.8.

3.2 Background: two time scales

In many service facilities customers are requested to make an appointment. There is a substantial body of literature focusing on the design of appointment systems. Healthcare is the most prevalent application area and hence also most considered in the literature (see the surveys by [104] and [267]). Appointment systems can be regarded as a combination of two distinct queueing systems. The first queueing system concerns customers making an appointment and waiting until the day the appointment takes place. The second queueing system concerns the process of a service session during a particular day. We denote these two queueing processes as

the 'access process' and the 'day process'. The remainder of this section provides an overview of the literature relevant for the present work and is structured as follows: (1) appointment scheduling, (2) access time models, and (3) integrating the access process and the day process.

3.2.1 Appointment scheduling

Appointment scheduling concerns designing blueprints for day-appointment schedules with typical objectives as minimizing customer waiting time, and maximizing resource utilization or minimizing resource idle time. A large part of the literature focuses on scheduling a given number of appointments on a particular day (e.g., [340, 396, 401, 595]). The extent to which various aspects that impact the performance of an appointment schedule are incorporated varies, such as customer punctuality (e.g., [390]), customers not showing up ('no-shows') (e.g., [308, 340]), lateness of the server at the start of a service session (e.g., [401]), service interruptions (e.g., [390]) and the variance of service duration (e.g., [308]).

Research techniques employed in appointment scheduling can be divided in analytical and simulation-based approaches, of which the latter is mostly applied [104]. In the day process we aim for an analytical approach, namely finite time Markov chain analysis. Related examples with healthcare applications are [291, 340, 396, 476, 595], although these references do not consider unscheduled customers.

Often, a homogeneous customer population is assumed [138]. Some studies however, focus on service systems with various customer types. Differentiation between customer types is identified as a consequence of distinct service requirements (e.g., [106, 351, 594, 595, 620]). Also, distinct priority levels may be a reason for patient type differentiation. An example can be found in [473], where service slots are premarked for various scheduled customer classes. In this paper, customer type differentiation arises from distinct arrival processes.

The effect of mixed arrival processes is studied in [257, 357, 532]. In these references, scheduled outpatients, unscheduled inpatients and emergency patients are taken into account. Patients without an appointment are either emergency patients who require non-preemptive priority or inpatients available for 'call-in' at any time during the day. These unscheduled patients are assumed to arrive according to an equal arrival rate throughout the day. In our case, we consider walk-in patients without priority who cannot be called in during the day. Moreover, we consider non-stationary arrivals to incorporate the expected peak behavior of walk-in demand. Studies that do incorporate non-priority unscheduled arrivals similar to the unscheduled arrivals in this paper are [20, 105, 106, 368, 497, 545, 548]; however, in all cases a simulation approach is employed. Also, these studies do not incorporate balking behavior of unscheduled customers.

3.2.2 Access time models

As our approach consists of a decomposition, isolated access time models are also of interest. The access process we consider is discrete-time and cyclical in both the

arrival and service processes. Various access time models based on continuous-time queueing models are available. Examples are the $M(t)|M|s(t)$ queue [258] and the adapted $M|M|s$ queue that models time-dependent demand [253]. The latter method is also applied to a healthcare problem in [259]. To preserve the discrete-time nature we take as starting point the generating function approach for slotted queueing models in discrete time [87]. A survey on discrete-time queueing systems is presented in [86].

Models to evaluate the length of hospital waiting lists are introduced in [641], and further studied in for example [238]. In these models homogeneous appointment request arrivals are assumed. In polling models, multiple queues are served by one server in cyclic order (see [551] for an overview). However, cyclic arrival rates and cyclic service capacity have not yet been incorporated in polling models.

3.2.3 Linking the access and the day process

We found only a few references that jointly consider the access and day process. In [496], the authors propose a two time scale model for the Emergency Department (ED) – Ward patient flow. The fast time scale of the ED is modeled by a continuous time Markov chain, while the slower time scale of the wards is modeled by a discrete time Markov chain. In [352] and [594], appointment schedules ranging over a horizon of several days are evaluated. The aim is to minimize the patient's waiting and the doctor's idle time, but the patient's access time is not studied in detail.

The advanced (or open) access methodology described by [449] also considers two time scales. With advanced access, a clinic leaves a fraction of appointment slots vacant for patients that request an appointment on the same day or within a couple of days. As many patients as possible are scheduled on the day they make an appointment request. One should determine the optimal ratio between the reserved capacity for long-term and same-day appointments [179]. This principle is slightly adapted in [402], where the demand for short term appointments is distributed over several days, to smooth the daily load of the system. The aim of the advanced access methodology is to minimize access time ("do today's work today"). Note that in an advanced access clinic patients do announce themselves in advance and make a (same-day) appointment, contrary to the type of unscheduled patients we consider, who just show up. Models that study the advanced access methodology usually focus on capacity distribution (e.g., [179, 490, 491]).

Formulating a model to design an appointment schedule considering two time scales is usually done using simulation techniques (e.g., [359]). An analytic approach is presented in [474], where the effect of capacity allocation among competing patient classes on access time targets is studied using techniques from Markov decision modeling and mathematical programming. An approach related to ours, although without the presence of walk-in patients, is given in [140]. The authors consider a service facility, and first develop a vacation queueing system to determine the access time. Subsequently, an appointment system is developed that calculates the waiting time at the facility.

3.3 Formal problem description

This section defines all modeling assumptions, defines the CAS, formally states the research goal and gives an overview of the proposed approach. Then, Sections 3.4 and 3.5 present two models to respectively evaluate the access time to the facility and the day schedule performance. In Section 3.6, the two models are connected by an algorithm, through which the best CAS is computed.

Assumptions. A facility consisting of R resources is operational during T time slots of length h, during each day in a cycle of D days. Two types of patients have to be served: scheduled and unscheduled patients. Service takes one time slot. Scheduled patients are given a specific date and time immediately when an appointment is requested. In addition, when the facility is temporarily congested, unscheduled patients are also offered an appointment: if the service of an unscheduled patient cannot start within g time slots after arrival, it will leave the facility and an appointment will be planned for another day. We will refer to such patients as *deferred* unscheduled patients, or just deferred patients. The first available appointment slot for scheduled and deferred patients is always the next day at the earliest. All appointments, both scheduled patients and deferred unscheduled patients, are scheduled according to a First-Come First-Served (FCFS) principle.

We assume a non-stationary Poisson process for the arrivals of appointment requests, with $\lambda^1, \ldots, \lambda^D$ the arrival rates for different days in the cycle. Next, during each day in the cycle, we assume a non-stationary Poisson arrival process for unscheduled patient arrivals, with slot-dependent arrival rates: χ_t^d for day $d = 1, \ldots, D$ and time slot $t = 1, \ldots, T$. Table 3.1 summarizes the notation introduced in this section.

Cyclic appointment schedule. To effectively counterbalance the non-stationarity at both the daily and cyclic (i.e., weekly, biweekly or monthly) level, we aim to design an appointment schedule that is cyclic. We introduce the CAS $C = (C^1, \ldots, C^D)$, with $C^d = (c_1^d, \ldots, c_T^d)$, where c_t^d specifies the maximum number of patients that may be scheduled in slot t on day d.

To find an adequate appointment schedule, we propose a decomposition. First, we introduce the concept of a capacity cycle $K = (k^1, \ldots, k^D)$, where k^d prescribes the maximum number of patients to schedule for day d. Second, given the capacity cycle K, the day plan is specified. In order to match the capacity cycle K, the day plan C^d should be such that $k^d = \sum_{t=1}^{T} c_t^d$.

Goal. An effective strategy balances the opportunities (1) for unscheduled patients to be served on the same day without long waiting time and (2) for scheduled patients to be served within an acceptable access time. To this end, we define the best policy as the cyclic appointment schedule in which the expected fraction of unscheduled patients served on the day of arrival, F, is maximized, while for scheduled patients the access time service level, $S(y)$, defined as the percentage of patients that is served within y days, is above a pre-specified norm $S^{norm}(y)$. The value of the vector $(y, S^{norm}(y))$ is chosen by the facility managers.

Table 3.1: Notation introduced in Section 3.3.

Symbol	*Description*
R	Number of resources
T	Number of time slots during a day
t	Time slot index ($t = 1, \ldots, T$)
h	Length of a time slot
D	Cycle length in days
d	Day index ($d = 1, \ldots, D$)
g	Patience of an unscheduled patient, expressed in the number of slots a patient is willing to wait
λ^d	Initial appointment request arrival rate on day d
χ_t^d	Unscheduled patient arrival rate on day d during time interval $(t-1, t]$
c_t^d	Maximum number of appointments to schedule in slot t on day d
C^d	Appointment schedule on day d, $C^d = (c_1^d, \ldots, c_T^d)$
C	Cyclic appointment schedule, $C = (C^1, \ldots, C^D)$
k^d	Maximum number of appointments to schedule on day d
K	Capacity cycle, $K = (k^1, \ldots, k^D)$
F	$\mathbb{E}$[Fraction of unscheduled patients to serve at day of arrival during one cycle]
$S(y)$	Access time service level: fraction of patients with access time not greater than y
$(y, S^{norm}(y))$	Access time service level requirement: fraction of patients with access time not greater than y is at least $S(y)$
ϕ^d	Distribution of the number of deferred patients on day d
γ^d	Total appointment request arrival distribution on day d
v^d	Expected number of deferred patients on day d

Approach. The best CAS is determined by employing an iterative algorithm that effectively utilizes our decomposition of the CAS in the capacity cycle and the day plan. Figure 3.1 provides an overview of the algorithm.

In each iteration, first, capacity cycles are generated with at most $R \cdot T$ appointments per day, for which the access time service level norm will be satisfied. All patients requesting an appointment are taken into account – thus both scheduled patients and deferred unscheduled patients. We derive the distribution of the number of deferred unscheduled patients ϕ^d, so that the distribution of the total number of appointment requests on day d is the sum of a Poisson distribution with parameter λ^d and the distribution ϕ^d. To assess whether specific capacity cycles with arrival distribution γ^d satisfy the access time norm $S^{norm}(y)$, a cyclic slotted queueing model is proposed (Model I, presented in Section 3.4).

Next, for each capacity cycle generated in the first step, the best day schedule is determined. Given the queue length probabilities resulting from Model I and the unscheduled patient arrival rates, χ_t^d, for each day the k^d appointments are distributed over the T time slots, such that the number of deferred unscheduled patients is minimized. To achieve this, a Markov reward model is presented (Model II, Section 3.5), which is used to calculate the performance of a specific day schedule.

Then, the capacity cycle that achieves the lowest expected number of deferred unscheduled patients over the entire cycle is chosen as the best cycle. If the expected

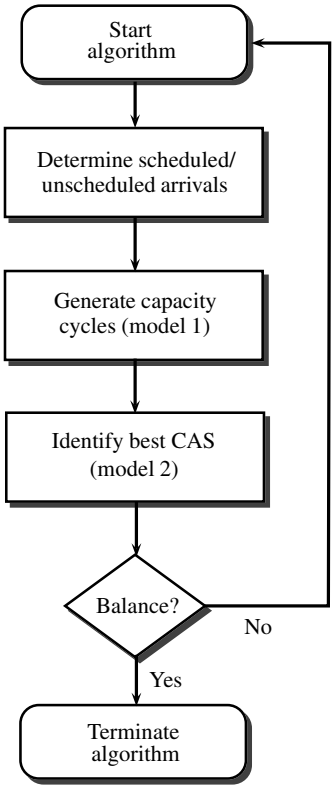


Figure 3.1: A flowchart of the algorithm.

numbers of deferred unscheduled patients v^d did not change significantly since the last iteration, the algorithm stops. If not, the entire process is repeated. A detailed description of the algorithm is given in Section 3.6.

3.4 Access time evaluation

In this section, a cyclic slotted queueing model (Model I) is presented to evaluate the access time for scheduled patients, given an arbitrary capacity cycle. To this purpose, we focus on the backlog, B^d, at the start of each day d. We define the backlog as the number of patients for which a request for an appointment has already been made, while the appointment itself has not yet taken place. We formulate a Lindley-type equation to characterize the backlog, and use a probability-generating function approach to derive expressions for the distribution of the backlog at the start of each day in the cycle. From the backlog distribution, we derive the access time distribution. A summary of the notation used in this section is given in Table 3.2.

Table 3.2: Notation introduced in Section 3.4 .

Symbol	*Description*
B^d	Backlog at start of day d
$P_{B^d}(z)$	Generating function of B^d
A^d	Number of appointment requests arriving at day d
a_j^d	Appointment request arrival probabilities, $P[A^d = j]$
$P_{A^d}(z)$	Generating function of A^d
π_j^d	Stationary backlog probabilities, $P[B^d = j]$
k	Total number of available appointment slots in a capacity cycle, $k = \sum_d k^d$
$\mathbb{E}[W^d]$	$\mathbb{E}$[Access time for an appointment request arriving at day d]
$\mathbb{E}[W]$	$\mathbb{E}$[Access time for an arbitrary appointment request]

3.4.1 Backlog distribution

Lindley-type equation. Consider day d. During the day, a maximum number of patients, k^d, is served, and a number of new patients, A^d, arrives. At the start of day d, there is a backlog B^d. Since it is not possible to make an appointment on the day of arrival itself, the backlog at the start of the next day equals the backlog on day d minus the number of patients served on day d plus the number of patients that arrived on day d. This can be formalized in the following Lindley-type equation:

$$B^{d+1} \quad = \quad (B^d - k^d)^+ + A^d,$$

where $(x)^+ = x$ if $x > 0$, and 0 otherwise.

Probability-generating function. Using an approach based on generating functions [87], we derive expressions for the distribution of the backlog at the start of each day in the cycle. The transition probabilities for going from state $B^d = i$ to state $B^{d+1} = i'$ are given by:

$$P[B^{d+1} = i' | B^d = i] = \begin{cases} P[A^d = i'] & \text{, if } i - k^d \leq 0, \\ P[A^d = i' - i + k^d] & \text{, if } i - k^d > 0. \end{cases}$$

We denote the stationary probability that at the start of day d, the backlog equals j patients by π_j^d. Furthermore, let a_j^d denote the probability that $A^d = j$. Note that the underlying probability distribution does not necessarily has to be Poisson. The stationary probabilities can be computed recursively, under the condition that the capacity for scheduled patients is larger than the average demand, i.e., $\sum_d \mathbb{E}[A^d] < \sum_d k^d$, since otherwise we would be dealing with an unstable system. For $d = 1, \ldots, D$, $j \geq 0$ we obtain:

$$\pi_j^{d+1} \quad = \quad a_j^d \sum_{i=0}^{k^d-1} \pi_i^d + \sum_{q=0}^{j} a_{j-q}^d \pi_{k^d+q}^d \, . \tag{3.1}$$

We multiply both sides of (3.1) with the complex number z^j, where $|z| \leq 1$, and z^j denotes z raised to the power j, as opposed to index d in π_j^d, a_j^d and k^d. The summation of both sides of the resulting equation over j yields the probability-generating

function for π^{d+1}:

$$P_{B^{d+1}}(z) = \sum_{j=0}^{\infty} \pi_j^{d+1} z^j = \sum_{j=0}^{\infty} \Big[a_j^d \sum_{i=0}^{k^d-1} \pi_i^d + \sum_{q=0}^{j} a_{j-q}^d \pi_{k^d+q}^d \Big] z^j.$$

From this we obtain:

$$P_{B^{d+1}}(z) = \sum_{j=0}^{\infty} \pi_j^{d+1} z^j = P_{A^d(z)} z^{-k^d} P_{B^d(z)} + P_{A^d(z)} z^{-k^d} \sum_{i=0}^{k^d-1} \pi_i^d \left(z^{k^d} - z^i \right).$$

Rearranging terms and changing the order of summation leads to the probability generating function of B^d:

$$P_{B^d}(z) = \frac{\sum_{i=1}^{D} \sum_{q=0}^{k^{d+D-i}-1} (z^{k^{d+D-i}} - z^q) \pi_q^{d+D-i} \left[\prod_{s=d}^{d+D-i-1} z^{k^s} \prod_{r=0}^{i-1} P_{A^{d+D-r-1}}(z) \right]}{\prod_{g=1}^{D} z^{k^g} - \prod_{h=1}^{D} P_{A^h}(z)},$$

where, since we consider days in a repeating cycle, we define:

$$d := \begin{cases} D & \text{, if } d \bmod D = 0, \\ d \bmod D & \text{, otherwise.} \end{cases}$$

The probability-generating functions uniquely determine the stationary probabilities $\pi_j^d, j = 0, \ldots, k^d - 1$, $d = 1, \ldots, D$. To calculate these probabilities, we build upon the approach given in [8]. Define k as the total number of available appointment slots in a capacity cycle, i.e., $k = \sum_{d=1}^{D} k^d$. Then, the denominator of $P_{B^d}(z)$ has $k-1$ zeros inside the unit disk; this can be shown by using Rouché's theorem [353]. All generating functions, including $P_{B^d}(z)$, are bounded for $|z| \leq 1$, and therefore the zeros of the denominator are also zeros of the numerator [87]. Thus we obtain $k-1$ equations, and use $P_{B^d}(1) = 1$ to secure the last equation. The $k-1$ zeros of the denominator of $P_{B^d}(z)$ can be found by solving:

$$\prod_{g=1}^{D} z^{k^g} - \prod_{h=1}^{D} P_{A^h}(z) = 0. \tag{3.2}$$

The solutions of (3.2) also represent zeros of the numerator. Together with the normalizing equation $P_{B^d}(1) = 1$, $P_{B^d}(z)$ is completely defined for $d = 1, \ldots, D$. Note that now only the backlog probabilities for $j = 0, \ldots, k^d - 1$, have been derived. The remaining backlog probabilities are calculated directly using (3.1).

3.4.2 Performance measures

The access time distribution can be directly derived from the backlog probabilities, since appointment requests are served according to the FCFS principle. The FCFS service order and the impossibility of making an appointment request for the day of arrival results in an access time of at least one day. Several performance measures

can be derived. Of particular interest are the probability distribution of the access time, the expected access time and the access time service level.

The probability distribution of the access time. First we derive the conditional access time probability that the access time for a client arriving at day d exceeds y days, given that the backlog at the start of day d equals b clients. As argued, for $y = 0$, we have

$$P[W^d > y | B^d = b] = 1 \quad \text{, for all } b.$$

For $y > 0$, we have:

$$P[W^d > y | B^d = b] = \begin{cases} 1 & \text{, if } b \geq \sum_{i=0}^{y} k^{d+i}, \\ \dfrac{\sum_{j=s+1}^{\infty} (j-s) \cdot P[A^d = j]}{\mathbb{E}[A^d]} & \text{, otherwise,} \end{cases} \tag{3.3}$$

where s represents the number of patients arrived on day d that will be served within y days:

$$s = \min\left\{ \sum_{i=1}^{y} k^{d+i}, \sum_{i=0}^{y} k^{d+i} - b \right\}.$$

We can explain formula (3.3) as follows. First, when the backlog b outnumbers the available capacity in y days, the conditional probability that the access time exceeds y days equals 1. Otherwise, all arrivals beyond the number s will wait for more than y days. There are $j - s$ such arrivals. Then, the probability that the access time for a client arriving at day d exceeds y days, equals

$$P[W^d > y] = \sum_{b=0}^{\infty} P[W^d > y | B^d = b] \cdot P[B^d = b].$$

The expected access time. Analogously, the expected access time for an appointment request that arrives on day d is computed with:

$$\mathbb{E}[W^d | B^d = b] \quad = \quad \sum_{y=0}^{\infty} P[W^d > y | B^d = b],$$

and thus

$$\mathbb{E}[W^d] \quad = \quad \sum_{b=0}^{\infty} \mathbb{E}[W^d | B^d = b] \cdot P[B^d = b],$$

and

$$\mathbb{E}[W] = \frac{1}{\sum_{d=1}^{D} \mathbb{E}[A^d]} \sum_{d=1}^{D} \mathbb{E}[W^d] \mathbb{E}[A^q].$$

The access time service level. Using the access time probability distribution, we determine the fraction of scheduled patients for which the access time does not exceed y. We define this as follows:

$$S(y) = \frac{1}{\sum_{d=1}^{D} \mathbb{E}[A^d]} \sum_{d=1}^{D} \Big(1 - P[W^d > y]\Big)\, \mathbb{E}[A^d].$$

3.5 Day process evaluation

This section presents the model (Model II) to evaluate the performance of a single day in the CAS. Recall that the CAS consists of a capacity cycle, $K = (k^1, \ldots, k^D)$, that prescribes the maximum number of patients that can be scheduled for day d. Using Model I, we were able to evaluate the access time performance of a given capacity cycle. Below, we evaluate the day process of a given appointment schedule, by formulating a Markov reward process.

3.5.1 Markov reward process

Note that although day appointment schedule C^d is open for scheduling appointments, there may be less backlog than the $k^d = \sum_t c_t^d$ available appointment slots. Therefore, we introduce the notation $\bar{C}^d$ to represent the *realized* day planning, which is the schedule we evaluate. Now, $\bar{C}^d = \left(\bar{c}_1^d, \ldots, \bar{c}_T^d\right)$ expresses the actually utilized appointment slots. Because appointments are planned on a FCFS basis, the realized appointment day schedule, $\bar{C}^d$, will always be a truncated version of the day schedule, C^d. The slots that are not utilized for appointments can be used for unscheduled patients.

Since we will consider the day performance on a day-by-day basis, in the remainder of this section we drop the superscript d for notational convenience. Table 3.3 provides a summary of the notation introduced in this section.

Assumptions. For clarity of presentation, some of the assumptions introduced in Section 3.3 are repeated. During one day the facility of R resources is operational during T intervals of length h. Two types of patients have to be served: scheduled and unscheduled patients. Service always takes one time slot of length h. At the beginning of each time slot, a service can start. If there are both scheduled and unscheduled patients, scheduled patients are given priority. Overtime is not allowed.

Scheduled patients arrive on time, according to the schedule $\bar{C}$. In addition, we allow for no-shows, that is, the probability that a scheduled patient actually arrives at the facility equals $1 - q$, so that q represents the probability that a patient does not show up.

Unscheduled patients arrive at the facility according to an inhomogeneous Poisson process with slot-dependent arrival rate χ_t. If the service of an unscheduled patient cannot start within g time slots after arriving, it will leave the facility and an appointment will be planned for another day. We assume that the facility has no

Table 3.3: Notation introduced in Section 3.5.

Symbol	*Description*
$\bar{C}$	Realized schedule under CAS C, $\bar{C} = (\bar{c}^1, \ldots, \bar{c}^D), \bar{c}^d = (\bar{c}_1^d, \ldots, \bar{c}_T^d)$
q	P[No-show of a scheduled patient]
e_t	Number of slots available for unscheduled patients in the next g intervals after time t
$p_t^s(s)$	P[Number of scheduled patients arriving at the start of slot $t = s$]
$p_t^u(u)$	P[Number of unscheduled patients arriving during interval $(t-1, t] = u$]
$P[(s,u)_{t+1} \mid (k,l)_t]$	Transition probability from state (t,k,l) to state $(t+1,s,u)$
$Q_t(s,u)$	P[Number of scheduled, unscheduled patients waiting at the start of slot $t = s,u$]
ν_t	$\mathbb{E}$[Number of deferred patients in time interval $(0,t]$]
ν	$\mathbb{E}$[Total number of deferred patients]
ϕ_t	Distribution of the number of deferred patients in time interval $(t-1,t]$
ϕ	Distribution of the total number of deferred patients

foreknowledge about potential no-shows. Therefore, an unscheduled patient arriving during interval $(t-1,t]$ will stay if –and only if– the number of unscheduled patients already waiting is strictly smaller than the minimum number of service slots during the upcoming g intervals that are not utilized by scheduled patients. The number of time slots anticipated to be available for unscheduled patients during the upcoming g intervals is denoted by e_t:

$$e_t = \sum_{j=t}^{\min\{t+g-1,T\}} (R - \bar{c}_j).$$

States. The state of the system is denoted by the tuple (t,s,u), which specifies that at the beginning of time slot t, s scheduled and u unscheduled patients are present.

Transition probabilities. Let $p_t^s(s)$ denote the probability that s scheduled patients arrive at the beginning of time slot t. Since each no-show is assumed to occur independently, these probabilities are calculated as follows:

$$p_t^s(s) = \binom{\bar{c}_t}{s}(1-q)^s(q)^{\bar{c}_t - s} \quad \text{, for } 0 \le s \le \bar{c}_t.$$

Let $p_t^u(u)$ denote the probability that u unscheduled patients arrive during time interval $(t-1,t]$. As specified, $p_t^u(u)$ is Poisson distributed with slot dependent parameter χ_t. Note that χ_1 represents the arrival rate of unscheduled patients that arrive before the opening time of the facility. Furthermore, note that any distribution function p_t^u can be used in the day process evaluation model. Therefore, for Model I the assumption of a Poisson arrival process is not strictly required.

Let $P[(s,u)_{t+1} \mid (v,w)_t]$ denote the transition probability of jumping from state (t,v,w) to $(t+1,s,u)$. Below we specify these transition probabilities for all possible events. In Figure 3.2, the state space for an arbitrary time slot t is displayed in which

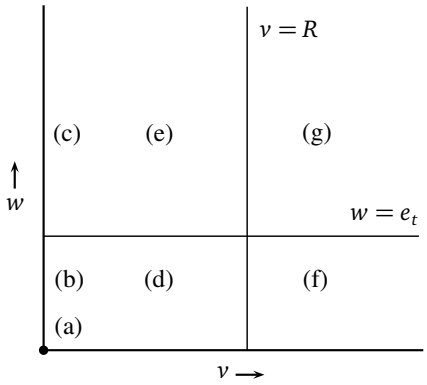


Figure 3.2: Day process state space and events.

the seven different possible events (a)-(g) are indicated. The events can be separated in three groups: first, cases (a)-(c) in which no scheduled patient is served ($v = 0$), second, cases (d) and (e) in which both scheduled and unscheduled patients are served ($v < R$), and third, cases (f) and (g) in which only scheduled patients are served ($v \geq R$). In the expressions below, $\mathbb{1}_A$ represents the indicator function; $\mathbb{1}_A = 1$ if condition A is satisfied, and 0 otherwise.

(a). $v = w = 0$; no patient served:

$$P[(s,u)_{t+1} \mid (v,w)_t] = p^s_{t+1}(s)p^u_{t+1}(u).$$

(b). $v = 0, 0 < w \leq e_t$; unscheduled patient(s) served:

$$P[(s,u)_{t+1} \mid (v,w)_t] = p^s_{t+1}(s)p^u_{t+1}(u - w + \min\{R,w\})\mathbb{1}_{(u \geq w - \min\{R,w\})}.$$

(c). $v = 0, w > e_t$; unscheduled patient(s) served and deferred:

$$P[(s,u)_{t+1} \mid (v,w)_t] = p^s_{t+1}(s)p^u_{t+1}(u - e_t + R)\mathbb{1}_{(u \geq e_t - R)}.$$

(d). $v < R, w \leq e_t$; scheduled and unscheduled patient(s) served:

$$P[(s,u)_{t+1} \mid (v,w)_t] = p^s_{t+1}(s)p^u_{t+1}(u - w + \min\{(R-v),w\})\mathbb{1}_{(u \geq w - \min\{(R-v),w\})}.$$

(e). $v < R, w > e_t$; scheduled and unscheduled served, unscheduled deferred:

$$P[(s,u)_{t+1} \mid (v,w)_t] = p^s_{t+1}(s)p^u_{t+1}(u - e_t + R - v)\mathbb{1}_{(u \geq e_t - R + v)}.$$

(f). $v \geq R, w \leq e_t$; scheduled patient(s) served:

$$P[(s,u)_{t+1} \mid (v,w)_t] = p^s_{t+1}(s - v + R)p^u_{t+1}(u - w)\mathbb{1}_{(s \geq v - R)}\mathbb{1}_{(u \geq w)}.$$

(g). $v \geq R, w > e_t$; scheduled patient(s) served, unscheduled patient(s) deferred:

$$P[(s,u)_{t+1} \mid (v,w)_t] = p^s_{t+1}(s - v + R)p^u_{t+1}(u - e_t)\mathbb{1}_{(s \geq v - R)}\mathbb{1}_{(u \geq e_t)}.$$

3.5.2 Performance measures

Let $Q_t(s,u)$ denote the probability that at the start of slot t there are s scheduled and u unscheduled patients present. $Q_t(s,u)$ can be calculated as follows:

$$Q_1(s,u) = p_1^s(s) \cdot p_1^u(u).$$

For $t = 2, ..., T$:

$$Q_{t+1}(s,u) = \sum_v \sum_w Q_t(v,w) P[(s,u)_{t+1} \mid (v,w)_t].$$

The expected number of deferred patients $v = v_T$ is calculated accordingly:

$$v_1 = \sum_{s=0}^{\infty} \sum_{u=e_1+1}^{\infty} (u - e_1) \cdot Q_1(s,u).$$

For $t = 2, ..., T$:

$$v_t = v_{t-1} + \sum_{s=0}^{\infty} \sum_{u=e_t+1}^{\infty} (u - e_t) \cdot Q_t(s,u).$$

The distribution of the number of deferred patients, ϕ, can be calculated as follows. For $t = 1, \ldots, T$:

$$\phi_t(j) = \begin{cases} \sum_{s=0}^{\infty} \sum_{u=0}^{e_t} Q_t(s,u) & \text{, if } j = 0, \\ \sum_{s=0}^{\infty} Q_t(s, e_t + j) & \text{, if } j > 0, \end{cases}$$

and

$$\phi = \phi_1 * \ldots * \phi_T,$$

where $*$ denotes the discrete convolution function.

Remark 3.1. Clearly, other performance measures that might be of interest, such as waiting time and utilization indicators, can also be calculated. Because in the algorithm of the next section, we will minimize the number of deferred patients, we restricted ourselves here to the calculation of this performance measure.

3.6 Algorithm

The algorithm presented in this section links the access and day process. Models I and II are used iteratively to maximize the number of unscheduled patients served during the day of arrival, given the specified access time service level norm. As mentioned before, unscheduled patients that cannot be served within g time slots receive an appointment. The algorithm determines the optimal size of this group

of deferred patients by gradually increasing its size during each iteration. Table 3.4 summarizes the notation presented in this section.

In the first iteration, the expected number of deferred patients is set to zero. Then, the best scheduling cycle (using Model I) with accompanying appointment schedule (using Model II) is determined, given the appointment request arrival processes with rate λ^d and that of unscheduled patient arrivals with rate χ_t^d. If the expected number of patients that has to be deferred under the best policy is significantly greater than in the previous iteration, then apparently the reserved capacity for appointments was not sufficient. In this case, the algorithm starts a new iteration. The distribution of the number of deferred patients on day d in iteration n is denoted by $\phi^d(n)$, and the expected number by $\nu^d(n)$.

In the subsequent iteration, to account for the patients that were deferred, the distribution of appointment request arrivals $\gamma^d(n)$ is set to

$$\gamma^d(n) = P(\lambda^d) * \phi^d(n-1),$$

where $P(\lambda^d)$ denotes the Poisson distribution with parameter λ^d. As such, the appointment requests generated by deferred patients are taken into account on the day of occurrence in the previous iteration. Then, a new best policy is calculated. As more appointment slots are reserved, this may result in more deferred patients than in the previous iteration. This iterative procedure is repeated until on each day in the cycle, a balance is found between the anticipated extra demand for appointments from deferred unscheduled patients (which was $\nu^d(n-1)$) and the realized deferred unscheduled patients (which is $\nu^d(n)$); expressed formally, the algorithm terminates if, for some small ϵ,

$$|\nu^d(n) - \nu^d(n-1)| < \epsilon \quad , d = 1, \ldots, D.$$

It is important to note that we aim for balance on a day-by-day basis. Balance just on a cycle basis ($|\sum_d \nu^d(n) - \nu^d(n-1)| < \epsilon$) is not sufficient, since only in the case

Table 3.4: Notation introduced in section 3.6.

Symbol	*Description*
n	Iteration counter
$\phi^d(n)$	Distribution of the number of deferred patients on day d in iteration n
$\nu^d(n)$	Expected number of deferred patients on day d in iteration n
$\gamma^d(n)$	Total appointment request arrival distribution on day d in iteration n
ϵ	Precision of the algorithm's stop criterion
$K(n_f)$	Capacity cycle option f consisting of $(k^1(n_f), \ldots, k^D(n_f))$ in iteration n
$C(n_f)$	The best CAS given capacity cycle $K(n_f)$
$\bar{\pi}_j^d(n_f)$	The probability that in iteration n under capacity cycle $K(n_f)$ j appointment reservations are utilized by appointments on day d
$\nu_C^*(n_f)$	$\mathbb{E}$[Total number of deferred patients in iteration n under capacity cycle $K(n_f)$ and CAS C]
$\nu^d_{C^d\|j}(n_f)$	$\mathbb{E}$[Number of deferred patients on day d in iteration n under capacity cycle $K(n_f)$ and CAS C when j appointment slots are utilized by scheduled patients]

that $|\nu^d(n) - \nu^d(n-1)| < \epsilon$, $d = 1, \ldots, D$, it is guaranteed that the appointment requests of deferred patients occur in the way that was anticipated. Only then we can assure that in the access time calculations, we account for the deferred patients on the day they occur, since the access time calculations that use $\phi^d(n-1)$, based on which the capacity cycle is designed, are still valid for $\phi^d(n)$ in this case.

Let us now specify the procedure emplyed to find an optimal policy within each iteration. First, by applying Model I, all capacity cycles fulfilling the specified access time service level norm are generated. So, given $\gamma^d(n)$, all capacity cycles $K = (k^1, \ldots, k^D)$ satisfying $S^{norm}(y)$ are generated. Suppose that m different capacity cycles satisfy the norm, then denote these options for iteration n by $K(n_f) = (k^1(n_f), \ldots, k^D(n_f))$, $f = 1, \ldots, m$. From these options, the best capacity cycle is selected, which is the cycle that minimizes the expected number of deferred patients. To do this, for each scheduling cycle option $K(n_f)$, the best CAS $C(n_f)$ is determined.

The best CAS's are determined by applying Model II as follows. First, observe that although in a capacity cycle $K(n_f)$ there are $k^d(n_f)$ appointment slots reserved on day d, not all of these reserved slots are necessarily utilized by scheduled patients. Since appointments are planned according to the FCFS principle, we know from the queue length probability vectors $\pi^d(n_f)$ of Model I, the probabilities of utilizing the first j out of the $k^d(n_f)$ reservations under capacity cycle $K(n_f)$. Let us denote these probabilities by $\bar{\pi}^d_j(n_f)$:

$$\bar{\pi}^d_j(n_f) = \begin{cases} \pi^d_j(n_f) & \text{, if } j = 0, \ldots, k^d(n_f) - 1, \\ \sum_{q=k^d(n_f)}^{\infty} \pi^d_q(n_f) & \text{, if } j = k^d(n_f). \end{cases}$$

By evaluating each day appointment schedule for $d = 1, \ldots, D$, $f = 1, \ldots, m$ and $j = 0, \ldots, k^d(n_f)$, the best CAS is determined for each capacity cycle $K(n_f)$, so by complete enumeration. Denote the expected total number of deferred patients in cycle $K(n_f)$ under appointment schedule C by $\nu_C(n_f)$. With $\nu^*(n_f)$ defined as the expected total number of deferred patients in cycle $K(n_f)$, under the best CAS the best cyclic appointment schedules are those that minimize:

$$\nu^*(n_f) = \min_C \nu_C(n_f) = \min_C \sum_{d=1}^{D} \sum_{j=0}^{k^d(n_f)} \bar{\pi}^d_j(n_f)\, \nu^d_{C^d|j}(n_f),$$

where $\nu^d_{C^d|j}(n_f)$ denotes the expected number of deferred patients on day d under capacity cycle $K(n_f)$ and cyclic appointment schedule C, if j appointment slots are utilized by scheduled patients. Note that $C^d|j$ is a truncated version of C^d, in exactly the same way that $\bar{C}^d$ was defined in Section 3.5. Now, the final step is to select the capacity cycle $K(n_f)$ and accompanying CAS, which is the CAS with the lowest expected number of deferred patients, namely:

$$\nu^*(n) = \min_f \nu^*(n_f), \;\; f^*(n) = \arg\min_f \nu^*(n_f), \;\; C^*(n) = \arg\min_C \nu_C(n_{f^*}).$$

Figure 3.3 displays the complete algorithm in pseudocode.

Step 1: *specify input*	*Specify:* $R, T, D, g, q, S^{norm}(y), \epsilon$; $\forall d : \lambda^d$; $\forall d, t : \chi_t^d$.
Step 2: *initialize algorithm*	$n := 1$; $\forall d : \nu^d(1) := 0, \gamma^d(1) := P(\lambda^d)$.
Step 3: *determine feasible cycles*	*Given* $\gamma^d(n)$, *determine all* $K(n_f), f = 1, \ldots, m$, *such that* $S(y) \geq S^{norm}(y)$. $\forall f, d$: *store* $\pi^d(n_f)$.
Step 4: *choose best cycle*	*Determine* $\nu^*(n), f^*(n)$ *and* C^*.
Step 5: *assess current solution*	*If* $\forall d : \lvert\nu^d(n) - \nu^d(n-1)\rvert < \epsilon$, *then stop, else proceed to step 6.*
Step 6: *adjust deferrals*	$\forall d : \nu^d(n+1) := \nu^d(n)$, $\phi^d(n+1) := \phi^d(n)$, $\gamma^d(n+1) := P(\lambda^d) * \phi^d(n+1)$; $n := n+1$ *and return to step 3.*

Figure 3.3: The algorithm in pseudocode.

Remark 3.2 (Convergence)**.** For the system to be stable we require that $\sum_d \lambda^d + \sum_d \sum_t \chi_t^d < R \cdot T$, so that total demand does not exceed capacity. In addition, we would like to determine the conditions under which the algorithm will converge. Therefore, first observe that since the unscheduled patient arrival rate χ_t^d is fixed and the first iteration starts with no deferred patients, i.e., $\nu^d(0) = 0$, in each iteration it is not possible to choose the CAS such that $\sum_d \nu^d(n) < \sum_d \nu^d(n-1)$. The total expected number of deferred patients $\sum_d \nu^d(n)$ is thus monotonically non-decreasing. Also, if the access time norm $S^{norm}(y)$ is set such that it can be satisfied if all patients are planned, we ensure that in each iteration it is possible to find feasible capacity cycles, i.e., capacity cycles for which $S(y) \geq S^{norm}(y)$. However, convergence of the algorithm is not assured. Although not likely for practical instances, it cannot be guaranteed that the algorithm does not run into the situation that it keeps jumping between points for which the total expected number of deferred patients does not change, but without day-by-day balance, i.e., $\left|\sum_d \nu^d(n) - \nu^d(n-1)\right| < \epsilon$, and not $|\nu^d(n) - \nu^d(n-1)| < \epsilon$, for all d. If such a case occurs, an additional rule to act as a tie-breaker is required. We extensively tested the algorithm by evaluating fifteen different instances (see Section 3.7). Convergence was obtained for all instances, also in the cases for which we tried to force the jumping behavior.

3.7 Numerical results

The algorithm was coded with the CodeGear Delphi programming language. We tested the algorithm on a variety of fifteen scenarios, each with different characteristics. To demonstrate our methodology, we choose to present one of the numerical

experiments in this section. First, we present the input parameters. Second, we discuss the evolution of the algorithm, and finally, we show the end results for the case study.

Input parameters. We consider a facility with one resource that employs a cycle with length $D = 5$ days, where each day consists of $T = 8$ slots. The initial demand per day for appointment requests is given by $(\lambda^1, \ldots, \lambda^5) = (5, 0, 2, 0, 7)$. The arrival rates of unscheduled patients χ_t^d are given in Table 3.5. These arrival rates are chosen such that different days in the cycle represent different unscheduled arrival patterns, as also illustrated by Figure 3.4. The access time service level norm is set such that 95% of the patients that are eventually scheduled are served within two cycles or less, $(y, S^{norm}(y)) = (10, 0.95)$. Furthermore, we assume that unscheduled patients are willing to wait for a maximum of two time slots, i.e., $g = 2$, and for computational convenience we assume that the number of deferred patients on day d, ϕ^d, is Poisson distributed. We assume that all scheduled patients show up, i.e., $q = 0$. The stop criterion of the algorithm applies the threshold $\epsilon = 0.0001$. Table 3.6 provides an overview of the input parameters. Note that the total expected demand for scheduled patients per cycle is 14, and the total expected demand for unscheduled patients per cycle is 22. Since there are $D \cdot T = 40$ time slots available within a cycle, the utilization of the system is 90%.

Execution of the algorithm. The algorithm was executed and the results obtained from each iteration are displayed in Table 3.7. In the first iteration the number of deferred unscheduled patients is positive on each day of the cycle, $v^d(1) > 0, d = 1, \ldots, D$. The total number of deferred patients is $\sum_d v^d(1) = 4.055$. Therefore, the deferred patients are added to the scheduled arrival stream and a new iteration is started. This procedure is repeated until after iteration 14, balance is obtained for each day, i.e., $|v^d(n) - v^d(n-1)| < \epsilon, d = 1, \ldots, D$. From Figure 3.5 and 3.6 it is seen that (as described in Remark 3.2, Section 3.6) the total number of deferred patients is monotonically non-decreasing, while deferrals on the day level are both increasing and decreasing. The fluctuations are substantial in the first iterations and the system stabilizes already after six iterations.

This behavior is also reflected by the dynamics of the capacity cycles found. The total number of reserved slots for appointment slots develops as follows: (16, 19, 21,

Table 3.5: Unscheduled patient arrival rates per slot per day.

χ_t^d	t								
d	1	2	3	4	5	6	7	8	Total
1	0.30	0.60	1.00	1.40	1.40	1.00	0.55	0.25	6.50
2	1.10	1.00	0.90	0.80	0.70	0.60	0.50	0.40	6.00
3	0.15	0.30	0.45	0.60	0.60	0.45	0.30	0.15	3.00
4	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.80
5	0.30	0.90	1.50	1.00	0.30	0.75	0.65	0.30	5.70

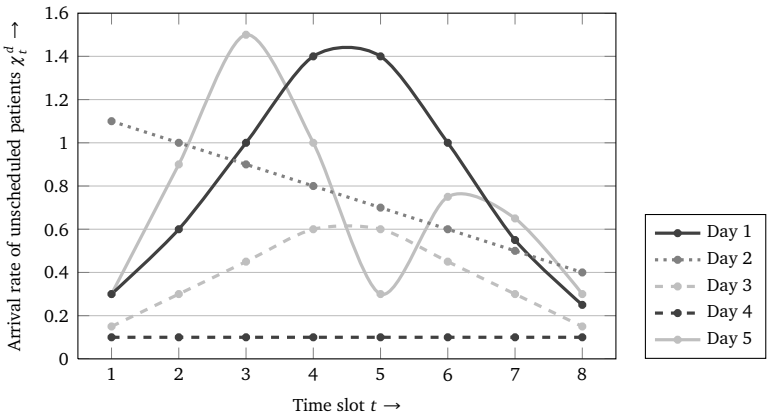


Figure 3.4: Graphical representation of the appointment request arrival rates per slot per day.

$21, 21, 22, \ldots, 22$). Again, although the total number of reserved slots $\sum_d k^d$ is monotonically non-decreasing, for a specific day k^d may also decrease. For example, the capacity cycles of iterations 3–5 all have a total capacity of 21, but the capacity cycle obtained in the third iteration is changed in iteration 4 so that one appointment is shifted from day 5 to day 3. This change is reversed in iteration 5. The final capacity cycle is already obtained in iteration 6. The only purpose of iteration 7–14 is to obtain the desired balance in the daily deferrals. Note that this is a result of the magnitude of ϵ. If ϵ had been set larger, the algorithm would have stopped earlier.

Results. Table 7.2 presents the final results for the numerical example. The percentage of unscheduled patients served on the day of arrival is 69%, so $F = 0.69$. This fraction is composed by fractions $F^1, \ldots, F^D$ that differ from day to day ($F^d = (\sum_t \chi_t^d - v^d)/\sum_t \chi_t^d$). For example, since day 4 is a quiet day with respect to unscheduled patient arrivals, it is completely filled with appointments. Only if no appointment request is made in one of the reserved slots, an unscheduled patient can be served. Apparently, it pays off to serve on average only 7% of the unscheduled patients directly on day 4 in the cycle. This is a result of the fact that only 3.6%

Table 3.6: Overview of the input parameters.

Parameter	*Description*	*Value*
D	Cycle length	5
T	Number of time slots	8
$\lambda^1, \ldots, \lambda^5$	Appointment request arrival rates	$(5, 0, 2, 0, 7)$
$(y, S^{norm}(y))$	Service level norm	$(10, 0.95)$
g	Unscheduled patient patience	2
q	No-show probability	0
ϵ	Algorithm precision	0.0001

Table 3.7: Results per iteration step of the algorithm.

Iteration	*Day*	*App. req. rate*	*Deferral rate*			*Cap. cycle*	*CAS*
n	d	γ^d	$v^d(n-1)$	$v^d(n)$	*difference*	k^d	C^d
1	1	5	0	1.133	1.133	1	(1,0,0,0,0,0,0,0)
	2	0	0	0.865	0.865	1	(1,0,0,0,0,0,0,0)
	3	2	0	0.547	0.547	4	(1,1,0,1,0,0,1,0)
	4	0	0	0.637	0.637	8	(1,1,1,1,1,1,1,1)
	5	7	0	0.873	0.873	2	(1,1,0,0,0,0,0,0)
2	1	6.133	1.133	1.456	0.323	2	(1,1,0,0,0,0,0,0)
	2	0.865	0.865	1.296	0.431	2	(1,0,0,0,0,0,1,0)
	3	2.547	0.547	0.549	0.002	4	(1,1,0,1,0,0,1,0)
	4	0.637	0.637	0.736	0.099	8	(1,1,1,1,1,1,1,1)
	5	7.873	0.873	1.371	0.498	3	(1,1,0,0,0,0,1,0)
3	1	6.456	1.456	1.456	0.000	2	(1,1,0,0,0,0,0,0)
	2	1.296	1.296	1.296	0.000	2	(1,0,0,0,0,0,1,0)
	3	2.549	0.549	0.952	0.403	5	(1,1,1,0,0,1,0,1)
	4	0.736	0.736	0.715	0.021	8	(1,1,1,1,1,1,1,1)
	5	8.371	1.371	1.752	0.381	4	(1,1,0,0,0,1,1,0)
4	1	6.456	1.456	1.456	0.000	2	(1,1,0,0,0,0,0,0)
	2	1.296	1.296	1.296	0.000	2	(1,0,0,0,0,0,1,0)
	3	2.952	0.952	1.498	0.546	6	(1,1,1,0,1,0,1,1)
	4	0.715	0.715	0.742	0.027	8	(1,1,1,1,1,1,1,1)
	5	8.752	1.752	1.402	0.350	3	(1,1,0,0,0,0,1,0)
5	1	6.456	1.456	1.456	0.000	2	(1,1,0,0,0,0,0,0)
	2	1.296	1.296	1.296	0.000	2	(1,0,0,0,0,0,1,0)
	3	3.498	1.498	0.954	0.544	5	(1,1,1,0,0,1,0,1)
	4	0.742	0.742	0.771	0.029	8	(1,1,1,1,1,1,1,1)
	5	8.402	1.402	2.049	0.647	4	(1,1,0,0,1,0,1,0)
6	1	6.456	1.456	1.456	0.000	2	(1,1,0,0,0,0,0,0)
	2	1.296	1.296	1.296	0.000	2	(1,0,0,0,0,0,1,0)
	3	2.954	0.954	1.495	0.541	6	(1,1,1,0,1,0,1,1)
	4	0.771	0.771	0.721	0.050	8	(1,1,1,1,1,1,1,1)
	5	9.049	2.049	1.794	0.255	4	(1,1,0,0,0,1,1,0)
		⋮				⋮	
14	1	6.456	1.456	1.456	0.000	2	(1,1,0,0,0,0,0,0)
	2	1.296	1.296	1.296	0.000	2	(1,0,0,0,0,0,1,0)
	3	3.497	1.497	1.497	0.000	6	(1,1,1,0,1,0,1,1)
	4	0.743	0.743	0.743	0.000	8	(1,1,1,1,1,1,1,1)
	5	8.897	1.897	1.897	0.000	4	(1,1,0,0,0,1,1,0)

of the unscheduled patients arrive on day 4, and that accordingly appointments are preferably planned on this day. The deferred unscheduled patients stream per day and the expected number of unscheduled patients served on the day of arrival are displayed in Table 7.2, which also reflects that on day 4 a small amount of unscheduled patients is directly served but also relatively few patients are deferred. The realized service level $S(10) = 0.962$ is well above the defined service level norm of 0.95.

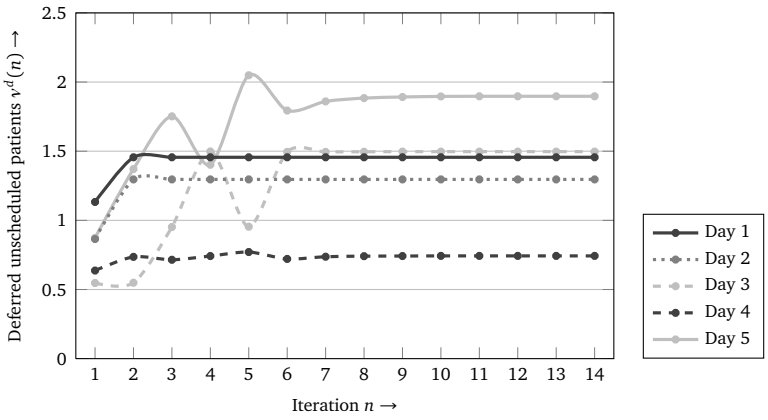


Figure 3.5: Graphical representation of the evolution of the deferral rates per day.

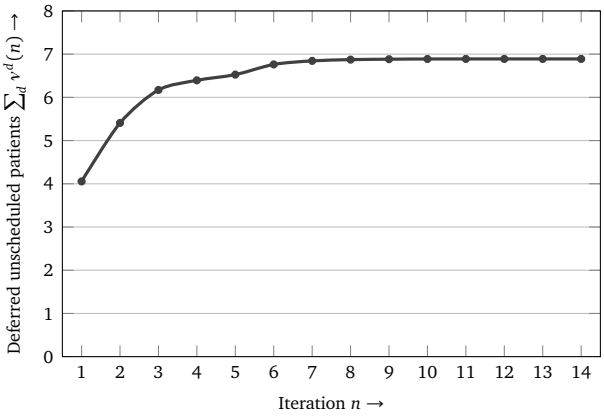


Figure 3.6: Graphical representation of the evolution of the total deferral rate.

The resulting capacity cycle is $K = (2, 2, 6, 8, 4)$, with corresponding day schedules which we discuss one-by-one below. Note that to achieve the service level norm it is required to reserve a buffer capacity of 1.11 to account for variability in appointment request arrivals, because 22 appointment slots are reserved while the average total number of patients to schedule within a cycle is $\sum_d (\lambda^d + v^d) = 14 + 6.89 = 20.89$. Apparently, the service level norm is achieved with only 5% buffer capacity, thus reserved capacity for appointments can be used efficiently.

The realized expected load per day, denoted by $L^1, \ldots, L^D$, is a result of the capacity cycle, the probabilities that the reserved appointment slots are utilized by appointment requests and the expected number of unscheduled patients served on day of arrival $\sum_t \chi_t^d - v^d$. It turns out that the load is balanced throughout the cycle where each day has a realized load between 6.7 and 7.7.

Finally, we discuss the resulting day schedules, to explain the moments on which the appointments are planned (see also Figure 3.7).

Day 1, $C^1 = (1,1,0,0,0,0,0,0)$**.** Although the lowest unscheduled arrival rate does occur at end of the day, the appointments are planned at the beginning of the day. Since unscheduled patients are willing to wait 2 time slots, a peak in arrivals has an impact until two slots afterwards. If appointments were planned at the end of the day, there is no possibility to serve arriving unscheduled patients, while when planning appointments at slots at the beginning of the day, early unscheduled arrivals can be served in the third time slot.

Day 2, $C^2 = (1,0,0,0,0,0,1,0)$**.** Again, the tendency to plan appointments early shows up. But, the drop in unscheduled arrivals is such that it is worthwhile to plan one appointment at the end of the day. However, again although the lowest arrival rate occurs in the latest time slot, the appointment is planned one slot before, to be able to serve an unscheduled patient arriving during interval $(T-3, T-1]$.

Day 3, $C^3 = (1,1,1,0,1,0,1,1)$**.** The demand for unscheduled patients is relatively low. Therefore, only two slots are left open in which no appointment is planned. These are planned during the peak hours of unscheduled arrivals. However, the open slots are not planned consecutively, so to spread the possibilities for unscheduled patient service.

Day 4, $C^4 = (1,1,1,1,1,1,1,1)$**.** As described, this is a quiet day for unscheduled patients. Therefore, all slots are reserved for scheduled patients. However, note that not always are all reserved slots used for appointments; 88% of the reserved slots on day 4 are utilized for scheduled patients.

Day 5, $C^4 = (1,1,0,0,0,1,1,0)$**.** The appointments are planned around the peaks of unscheduled arrivals. It is remarkable that the two later appointments do not

Table 3.8: End results for the case study.

Indicator	*Description*	*Value*
F	Fraction unscheduled directly served	0.69
$F^1, \ldots, F^5$	Daily fraction unscheduled directly served	$0.78, 0.78, 0.50, 0.07, 0.67$
$S(10)$	Service level scheduled patients	0.962
$v^1, \ldots, v^D$	Deferral rate per day	$1.46, 1.30, 1.50, 0.74, 1.90$
$\sum_t \chi_t^1 - v^1, \ldots, \sum_t \chi_t^D - v^D$	Unscheduled patient service rate per day	$5.04, 4.70, 1.50, 0.06, 3.80$
$L^1, \ldots, L^D$	Realized utilization per day	$7.04, 6.70, 7.48, 7.71, 7.06$
K	Capacity cycle	$(2,2,6,8,4)$
C^1	CAS day 1	$(1,1,0,0,0,0,0,0)$
C^2	CAS day 2	$(1,0,0,0,0,0,1,0)$
C^3	CAS day 3	$(1,1,1,0,1,0,1,1)$
C^4	CAS day 4	$(1,1,1,1,1,1,1,1)$
C^5	CAS day 5	$(1,1,0,0,0,1,1,0)$

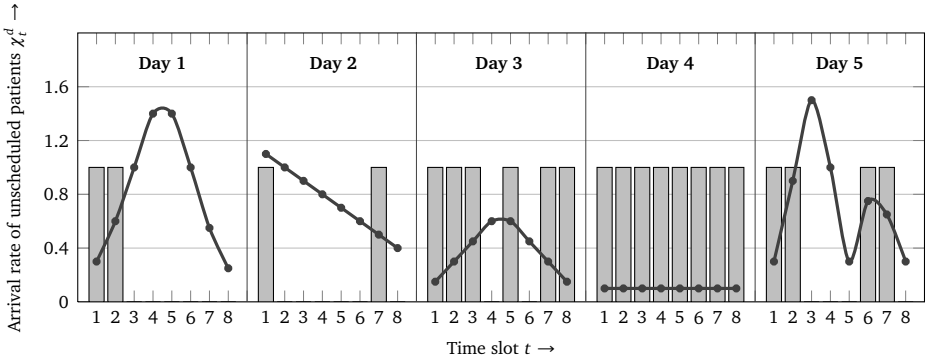


Figure 3.7: The CAS (bars) versus the unscheduled patient arrival rates (lines).

occur exactly during the off-peak hours but later, which can also be explained by the aforementioned delayed impact of unscheduled arrival peaks.

The final conclusion is that the resulting CAS and its performance is the outcome of the complex interaction between the scheduled patient arrival rates λ^d, the unscheduled patients arrival patterns χ_t^d, and the service level requirement $S^{norm}(y)$. For example, if $S^{norm}(y)$ is set tighter, the resulting capacity cycles more closely resemble the total arrival rates for appointment requests γ^d. Also, since there is less flexibility to spread the appointments, in that case the fraction of unscheduled patients served on the day of arrival, F, decreases.

3.8 Discussion

In this chapter, we have outlined a methodology to develop appointment schedule blueprints for facilities with scheduled and unscheduled arrival streams. The methodology consists of two separate models, one to evaluate the access and the other to evaluate the day process. The two models are linked by an iterative algorithm. An advantage of this modular approach is that the models and the algorithm can be updated separately, so that a high level of flexibility is obtained.

The chapter focused on developing a methodology that incorporates the key characteristics of a mixed system and an effective communication between the two time scales of the access and day process. Achieving numerical efficiency will be our next challenge. For the problem instance in Section 3.7, the CAS was found using complete enumeration. Our work is currently aimed at incorporating heuristics so that larger, more realistic instances can be evaluated. The model structure of the day process suggests that local search techniques are worth exploring (see for example [340, 595, 594]).

Some extensions can readily be incorporated in our approach. Management is free to choose the service level norm for the access time. As such, the resulting appointment schedules can be compared for several service levels. Also, different

choices for the time patients are willing to wait ('patient patience') or overbooking to anticipate for no-shows could be studied. Furthermore, the access time for scheduled patients and the fraction of unscheduled patients who cannot be served on the day of arrival are outcomes of Model I and Model II respectively, and serve as input for the algorithm. Of course, other model outcomes could be chosen as well. Finally, to incorporate for example planned maintenance of a service facility, the number of available slots in the day process can easily be amended by closing slots. Worthwhile to consider would also be to introduce stochastic service times and stochastic patient patience in the day process. Last but not least, our focus will be on practical issues in the implementation of the methodology at outpatient care facilities that (will) allow walk-in; to begin with, at the Academic Medical Center Amsterdam.

CHAPTER 4

Organizing Multidisciplinary Focused Care Facilities

4.1 Introduction

The Academic Medical Center (AMC) Amsterdam, opened in 2011 a center for children with neuromuscular diseases. The purpose of this center, the 'Children's Muscle Center Amsterdam' (CMCA), is to enhance the quality of care, by improving the coordination of care, clustering the expertise of the involved care providers, and drastically reducing the required number of hospital visits. This chapter describes how quantitative modeling supports the AMC in the design and operations of the CMCA. This study provides an example of how techniques from Operations Research and Management Sciences (OR/MS) can be employed to help realizing cost-efficient care facilities that offer focused care to patients with specific complex diseases via the one-stop-shop principle.

4.1.1 Children's Muscle Center Amsterdam

Neuromuscular diseases is the generic term for a broad set of disorders which impair the functioning of the muscles via muscle or nerve pathology. Most of the diseases are progressive in time, sometimes leading to an early death of the patient [200]. Most neuromuscular diseases have no cure, so the goal of the treatment is to reduce symptoms, and increase both mobility and life expectancy [439]. Examples of neuromuscular diseases are the diseases of Duchenne, Becker and Charcot Marie Tooth [601]. Children with neuromuscular diseases typically need care from various physicians and therapists.

Due to their disease, children may suffer from a variety of symptoms. Therefore, children are usually seen by a rehabilitation physician, a neurologist, a clinical geneticist, a cardiologist, and a pneumonologist. In addition, psychologists, dietitians and even cardiac surgeons may be required. Accurate coordination of such multidisciplinary treatment is crucial to achieve high quality of care. If such coordination is insufficient, under- or overtreatment may take place, treatments may be performed in a non-optimal order, or certain aspects of the disease may be overlooked. However, since the different disciplines are accommodated at different locations within the hospital, this coordination is a challenging task.

With the opening of the CMCA, care coordination for children will be significantly improved. Regularly, a treatment day is organized on which the required disciplines come together to see multiple patients. Physicians discuss the condition of the different patients, so that diagnoses are settled earlier and treatments are better customized. Through the establishment of the CMCA, children and their parents will generally visit the hospital only once a year, while previously they visited on average six times a year. This is a major improvement, because the hospital visits are both physically and psychologically demanding for the patients. Also, the great responsibility and burden is taken away from the parents: to gather all relevant information from the different hospital visits and to schedule the right appointments at the right point in time. Consequently, the CMCA will simultaneously increase quality of care and patient-centeredness.

However, realizing centralized care is not a challenging task. First of all, all physicians should cooperate and reserve time for the treatment days in their already busy schedules. Second, for each treatment day patients have to be selected and scheduled in an optimal way. Since the schedules are heavily constrained, construction by hand is very time consuming and does not guarantee the best solution. Third, due to the small size of the patient group, the treatment days are not often organized, and due to the many constraints, only a few patients can be scheduled in one treatment day. Consequently long access times may arise. This chapter will show that OR/MS techniques can be very helpful in these design and control issues. It contributes to the logistic questions on two levels:

Treatment day scheduling. First, we present a day scheduling algorithm to address the challenging task of scheduling the required combinations of consultations, diagnostics and treatment in combination on one day. By the analysis of historical data and interviews with physicians and therapists, we collected all relevant restrictions and preferences. Based on this information, we developed an Integer Linear Program (ILP) in close cooperation with the neuromuscular disease care experts. This ILP simultaneously selects which patients to invite for a particular treatment day, and generates an optimal day schedule, in compliance with all restrictions and preferences.

Access time evaluation. Second, we present a model to derive the probability distribution of the access times of newly diagnosed patients. Computer simulations are performed in which the scheduling algorithm is iteratively applied. As such, probability distributions of the number of patients that can be invited to one treatment day is obtained. These distributions are input for a Markov model by which access time distributions are determined. Since the CMCA has started very recently demand predictions are very uncertain, estimations run from 20 to 50 new patients per year. Therefore, various demand scenarios are considered. The influence of the several constraints on the day schedules on access times are analyzed and improvements are proposed.

The outcomes of this study are used to advise the AMC on how often treatment days should be organized, for which care providers the availability should be reconsidered, and which day-schedule preferences could better be dropped. For example,

initially the physicians were to have at least two appointments per day. Our research showed that such a constraint would result in very long access times because only patients with the same needs could be scheduled together. Based on our results, this constraint and several other highly restricting constraints were adjusted or removed.

4.1.2 Literature

Recall from Chapter 3 that appointment systems can be regarded as a combination of two distinct queueing systems. The first queueing system concerns customers making an appointment and waiting until the day the appointment takes place. The second queueing system concerns the process of a service session during a particular day. The literature review [267] identifies several open challenges in appointment scheduling, of which prominent ones are planning coordinated packages of care for patients needing treatment from several health services, scheduling in highly constrained situations, and linking the access process and the day process. These challenges are addressed in the current chapter. Note that the addressed references in Section 3.2, which we do not repeat, are also relevant to the current chapter.

The literature has mostly focused on scheduling a given number of single appointments on a particular day for an individual service provider [104]. Scheduling multiple appointments at once for a single discipline for a planning horizon of one day or one week is done in [119, 120, 484], without considering access times. In these references, given sets of physical therapy treatments of given sets of patients on a particular day are scheduled. In [119, 120] a formulation of this problem as a hybrid shop scheduling problem is presented, which is solved by a genetic algorithm [120], combined with data mining techniques in later work [119]. A scheduling algorithm based on genetic algorithms and machine learning is described in [484].

ILP approaches for highly constrained *mono*disciplinary treatment planning can be found in [132, 135, 462, 572] for radiotherapy and chemotherapy treatment planning. For these patients treatments have to be scheduled during a given number of weeks, strictly taking into account the required rest periods. When access times have to be minimized, it is important to have good rules according to which patients are selected to be admitted from the waiting lists [337, 604]. An ILP for radiotherapy treatment planning is described in [132, 135], so that a maximum number of patients is planned from the waiting list, thereby minimizing the access time of patients while maximizing device utilization. However, they do not explicitly evaluate access times. In [572] a two stage ILP approach for solving a similar problem is presented, but without the patient selection decision. In the first stage patients are assigned to days, and in the second stage appointment times are given to all patients on their assigned days. The objectives are minimizing access times, treatment delays and staff overtime. A time horizon of a week is considered in [462], in which one appointment per patient should be planned for a single discipline. To resolve computational difficulty, in [462] an ILP model is proposed, that is broken down into three manageable hierarchical stages. In the first stage patients are selected, in the second stage patients are assigned to therapists, and in the third stage patients are

scheduled throughout a day.

To evaluate access times, Chapter 3 presented a slotted queueing model in discrete time that is solved by a generating function approach. The access time model formulated in the current chapter has the purpose to evaluate access times as waiting times of customers in a queue with batch service, where the batch size is *at least two*. The maximum batch size is derived from the solution of the appointment scheduling problem. Queueing systems with batch service were first considered by Bailey [24], motivated, as in this chapter, by evaluation of access times for out-patients in hospitals. Other applications mentioned in the literature are in transport, control of traffic flows, and manufacturing. There is a vast literature on the analysis and numerical evaluation of queues with batch service, see for example [116, 239, 454, 455]. In this chapter, we approximate the queueing process with a finite Markov chain and use the renewal theory to derive stationary waiting times.

The chapter is organized as follows. Section 4.2 describes the characteristics of the case study setting. Section 4.3 presents the ILP model for planning of a treatment day. The planning algorithm is applied to data of the aimed patient group, and the results for these patients are presented. Based on the results of this planning algorithm, an access time model is derived in Section 4.4, and numerical results are given for the AMC case. The chapter ends with a discussion in Section 4.5.

4.2 Background: case study

The CMCA aims at children up to eighteen years old who have a neuromuscular disease (follow-up patients), or are suspected of having one (new patients). They do invite follow-up and new patients on different days, because a different team of physicians is required. Therefore, 'diagnosis days' are organized for patients suspected of having a neuromuscular disease, and 'follow-up days' for patients who have already been diagnosed. Figure 4.1 displays an overview of the patient flow.

Preconsultation. When a physician suspects a neuromuscular disease, the patient and the parents are first asked to fill out a questionnaire at home, which is then assessed by the CMCA. If the questionnaire does not support the suspicion of a neuromuscular disease, the patient will not be invited to the CMCA.

Diagnosis day. If a patient is eligible for a diagnosis day, a set of required consultations and examinations is determined during a meeting between the 'core members' of the multidisciplinary treatment team. The core members are the physicians who are together responsible for the patient's treatment. They will all see the patient during the diagnosis day. The team is completed by a nurse practitioner who provides administrative support to both physicians and patients. For diagnosis days the core of the team is formed by a paediatric neurologist, a clinical geneticist, and the nurse practitioner.

Next, the patient is scheduled to come to the AMC for a diagnosis day. On this day, there will first be an intake meeting between the patient and the nurse practitioner. Then, the prescribed consultations and examinations will take place. Halfway

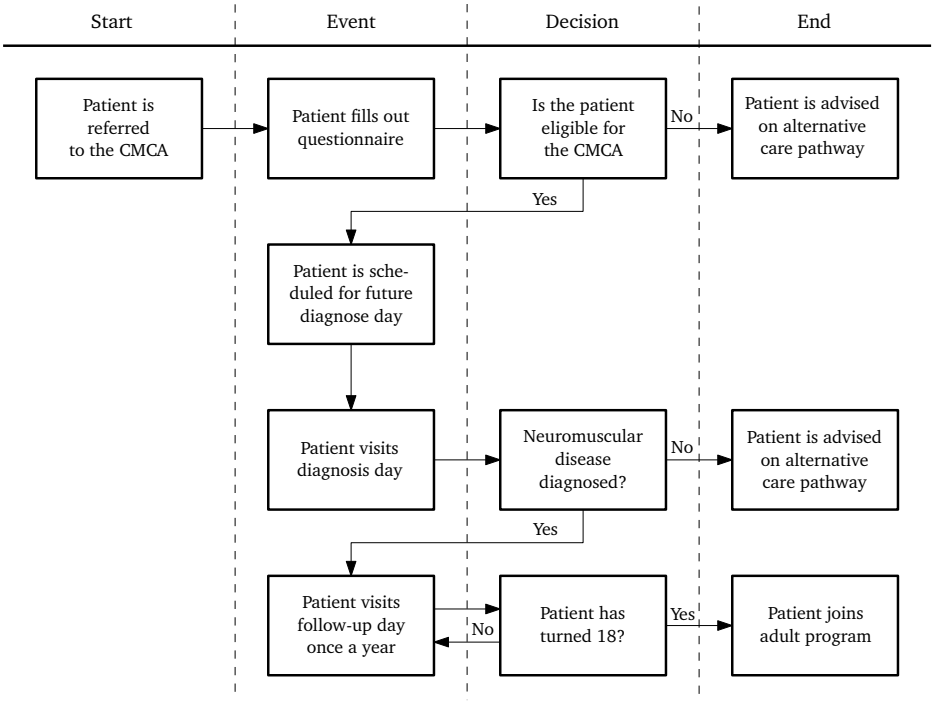


Figure 4.1: Patient flow diagram.

the afternoon, a Multidisciplinary Team Meeting (MTM) is scheduled in which the outcomes of the consultations and examinations are discussed. If possible, a diagnosis is settled, and a care plan is designed. Otherwise, additional examinations are scheduled (not on the same day). After the MTM, the neurologist shares the conclusions with the patient. During the day the nurse practitioner is present to act as a host for the patient and to guide the patient through the different examinations. The day finalizes with a meeting between the patient, parents, the nurse practitioner, to answer remaining questions and to explain the further care pathway.

Follow-up programme. If the conclusion of the diagnosis day is that the patient has a neuromuscular disease, he will continue to the follow-up programme. During this programme, the physicians monitor the health of the patient and give advice on how to reduce and handle symptoms. Most of the check-ups have to be performed annually, therefore the patient will visit a follow-up day once a year. Here, the core of the team consists of a paediatric neurologist, a paediatrician, a rehabilitation physician and the nurse practitioner. Also, the set of required appointments is different than for the diagnosis day and depends on the type and severity of the disease a patient suffers from. The set-up of a follow-up day is as follows: an intake with the nurse practitioner, examinations, a first MTM, consultations, a second MTM, a feedback consultation by a rehabilitation physician, and a final consultation with the nurse

Table 4.1: The required appointments per patient type type.

		Appointment (minutes)														
		Necessary				*Desirable*										
		Intake and evaluation (3 x 15)	Clinical geneticist (45)	Neurologist (45)	MTM (15)	Paediatrician (30)	Rehabilitation physician (30)	Physiotherapist (45)	Blood examination (15)	Clinical photograph (15)	Cardiac ultrasound / ECG (60)	EMG (75)	Muscle ultrasound (45)	MRI (60)	X-ray (30)	Needle biopsy muscle (120)
Patient type	*%*	*Required by x% of the patients*														
Myopathy (MP)	39	100	100	100	100	50	50	50	100	100	50	15	100	15	20	15
Neuropathy (NP)	39	100	100	100	100	50	50	-	100	100	-	50	-	-	-	-
Spinal muscular atrophy (SMA)	20	100	100	100	100	50	50	50	100	100	-	15	50	-	10	-
Neuromotor disease (NMD)	2	100	100	100	100	50	50	-	50	100	-	-	-	-	-	-

practitioner. When the patient turns 18, he will proceed to the adults track.

Day schedule. A month prior to a diagnosis day, the patients are selected who are preferably invited for the diagnosis day of the next month. If there are it at least two candidate patients, a diagnosis day is scheduled. Otherwise, the hospital considers it to be inefficient. A feasible day schedule has to be composed to asses how many patients can actually be invited. The schedule of a treatment day is highly constrained: some physicians are only available on specific times of the day, for some consultations several physicians have to be present, appointment precedence constraints have to be satisfied, all results of diagnostic tests have to be available before the MTM can start, etcetera. Based on the predominant suspected disease, patients are assigned to one out of four patient types. For each type, the percentage of patients that require a certain appointment is listed in Table 4.1. These numbers are based on estimations of the involved physicians and data from the patients who have been treated before, outside the CMCA.

The members of the core team subdivide for each patient the list of required consultations and examinations in 'necessary' and 'desirable' appointments. If all necessary appointments can be scheduled, a patient is invited to a diagnosis day. Further, the trade-off has to be addressed between skipping some of the desirable (but not necessary) appointments in order to invite more patients, or keeping all appointments and inviting less patients. Clearly, the latter option will result in longer access times. In Section 4.3, an integer linear program is presented that simultaneously addresses: (1) rational patient selection in conjunction with the appointments to be executed, and (2) the creation of a day schedule. The resulting access times are analyzed in Section 4.4.

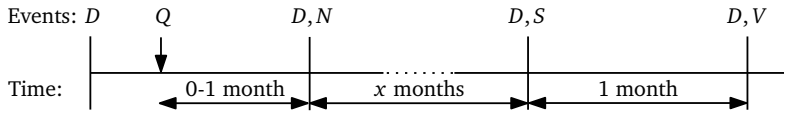


Figure 4.2: The time line of the patient access times to diagnosis days (Legend: $D =$ diagnosis day, $Q =$ questionnaire received, $N =$ not yet scheduled because of the waiting list, $S =$ scheduled for next treatment day, $V =$ visit takes place).

Access times. The next concern in designing and operating the CMCA are the access times that are to be expected for diagnostic patients. For follow-up days access times are no major issue, since the candidate patient are known well in advance, and patients are required to revisit between 12 to 15 months after their last treatment day. Therefore, the access times for follow-up patients are well-predictable. This, in contrast to the access times of new patients. For these patients the access time is counted as the number of days between the reception of the completed questionnaire and the patient's visit. The time line for a patient to get access to a diagnosis day is illustrated in Figure 4.2.

The AMC strives for a maximum access time of seven weeks for diagnosis days, which is quite ambitious in the initial setting, as diagnosis days are initially intended to be planned once a month. The access times will grow rapidly if the number of patients treated in a diagnosis day is insufficient. However, due to the variety in patient types, the complexity of the set of scheduling constraints, and diverging availability of the different care providers, the number of patients that can be invited for each diagnosis day cannot be easily predicted. Section 4.4 addresses this issue by presenting a Markov model by which the access time distribution for diagnosis patients is derived based on the probability distribution of the number of patients that can be seen on a particular diagnosis day.

4.3 Day schedules

In this section, a mathematical model is formulated and implemented to decide which patients are invited to visit the center for the next treatment day (both diagnosis and follow-up) and to compute an optimal schedule for this day. We first give an overview of the properties of the model. Next, we make some remarks on computation of the solution, and finally, present the results for diagnosis days. For clarity of presentation, the detailed mathematical formulation of the model is presented in the appendix. The model was developed in close cooperation with the CMCA healthcare professionals. Several versions of the model were designed and tested. Each time, the formulation and the inclusion or exclusion of specific constraints and objectives were discussed based on the outcomes on various test problems. Here, we restrict ourselves to presenting the end result.

4.3.1 Model formulation

We model the construction of a day schedule for a treatment day as an ILP. To do so, we divide a treatment day in time slots of equal length. The decisions to be made are: which patient gets which appointment at what time slot with which resource(s), which can be staff members and/or equipment. Thus, the decision variables are:

$$z_{j,p,s,t} = \begin{cases} 1 & \text{, if patient } j \text{ has an appointment with resource } s \text{ for} \\ & \text{procedure } p \text{ starting at time slot } t, \\ 0 & \text{, otherwise.} \end{cases} \tag{4.1}$$

We say that a patient has a *complete visit* if all his/her appointments are scheduled (both the necessary and desirable appointments, recall Section 4.2). If some of the desirable (but not necessary) appointments are omitted we say that a patient has a *partial visit*. We are interested in which patients have a complete visit, which have a partial visit, and at what time they have which appointment. The formal description of these variables can be found in Table 4.5, see Appendix 4.6.1. The constraints and objectives will be described in the next subsections, the mathematical formulation is given in Appendix 4.6.1.

Constraints. We distinguish several types of constraints:

Patient selection. A patient has most of his appointments, all of them, or none. The visit of such a patient is thus a complete visit, a partial visit, or the patient is not scheduled. The amount of appointments that are allowed to be omitted is patient specific. At least two patients should be scheduled on a treatment day for it to take place. Patients are scheduled according to the FCFS discipline.

Basic constraints. A patient gets each treatment at most once, a treatment is carried out by a resource that has the necessary qualifications, a resource can only be scheduled at one place at a time and should be available.

Precedence constraints. Some treatments have to be performed before others, and for some combinations of treatments there is a minimum amount of time in between the starting times of these treatment.

CMCA specific appointment constraints. Some of the appointments of a patient can take place simultaneously. For example, an orthopedist and physiotherapist can treat one patient at the same time. A patient needs time to rest, therefore, at least half an hour per three hours must be free from appointments. Some appointments are obligatory. If these are not scheduled, the patient cannot be scheduled.

Patient specific constraints. A patient cannot get more appointments than he or she can physically handle in a day. Therefore, sometimes appointments of a patient have to be spread over several days. This number of days is given per patient. However, an appropriate fraction of total appointment time should be scheduled on the first treatment day in order to avoid scheduling problems in the future.

MTM constraints. The Multidisciplinary Team Meeting is scheduled as a series of appointments, one per patient, in consecutive time slots, to make it just one meeting. All core team members should be present at the MTM. In some cases, the MTM must start at a fixed time. In other cases, this time may be flexible.

Defining constraints. The final constraints are required for the mathematical formulation. They determine the starting and end times of the patients and staff members, and determine whether a patient has an appointment at a certain time or not.

Objectives. The following objectives have been formulated, in descending order of priority:

1. Maximize the number of patients that have a complete visit
2. Maximize the number of patients that have a partial visit
3. Maximize the treatment time of all scheduled patients
4. Minimize the idle time in the schedules of the clinicians
5. Minimize the idle time in the schedules of the patients

The objective of the ILP is to maximize the sum of the weighted rewards on these objectives. In the objective function (see (4.15)), objectives 1, 2, 3, 4, and 5 are respectively rewarded by weight factors $\alpha, \beta, \gamma, \delta$, and ϵ. One may observe that the objective function contains multiple goals that are possibly in conflict. By varying the weight factors in the objective function, the relative importance of the various goals can be specified.

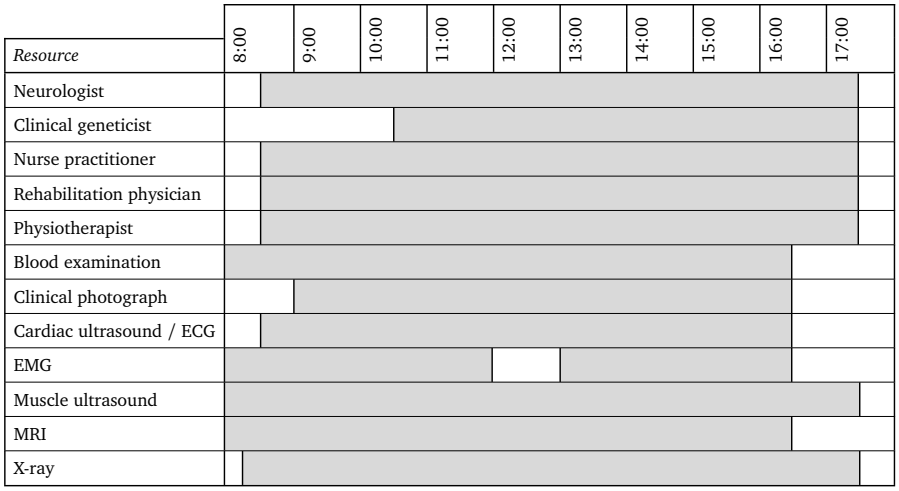
4.3.2 Computed schedules

The ILP coded with the program AIMMS. The solver employed is Cplex 12.2, using the branch and bound technique. A run is stopped as soon as the gap between the LP-bound and the best solution so far is less than 1%, thus, when a nearly optimal solution has been found. The first four objectives have always found their optimal value by then, the fifth not necessarily.

All input parameters of the ILP are set according to the CMCA data. Table 4.2 displays the availability of all resources during a diagnosis day. We choose to not include other data such as appointments precedence and qualifications of the staff to carry out the procedures here. Table 4.3 lists the values used for the weight factors. To determine these values, the CMCA clinicians scored the relative importance of objective on a 0–10 scale. As the objectives are not measures on the same scale, we applied a normalization factor to each factor in order to get comparable measures. These normalization factors, multiplied by the relative importance, resulted after several calibration runs in the listed weight factor values. An example of the result for the schedule of a diagnosis day is displayed in Figure 4.3.

Several bottlenecks are identified in the scheduling of diagnosis days. The following issues restrict the capacity of the CMCA, and need consideration when the CMCA desires to expand:

Table 4.2: Resource availabilities on a diagnosis day.



- Each patient has to visit the clinical geneticist for 45 minutes. However, this physician is available only from 10:30. These consultations have to take place before the examinations. Since a fifth patient can visit the clinical geneticist at the earliest at 13:30, just a little time is left for the examinations.
- The results of the blood examination have to be known before the MTM. However, obtaining these results takes two hours, and the blood examination cannot be done before the consultation with the clinical geneticist. Thus, at most three patients that can have a blood examination, regardless of their other appointments.
- Each patient has two appointments after the MTM, one with the neurologist and one with the nurse practitioner. In combination with the growing length of the MTM as there are more patients, this results in less time for consultations and examinations before the MTM.
- Half of the patients with a neurological disease need to have an EMG examination. The examination takes more than an hour, and the outpatient clinic is

Table 4.3: Weight factor values.

Objective	*Weight factor*	*Importance*	*Value*
1	α	10	100
2	β	8	50
3	γ	10	2
4	δ	6	20
5	ϵ	5	2

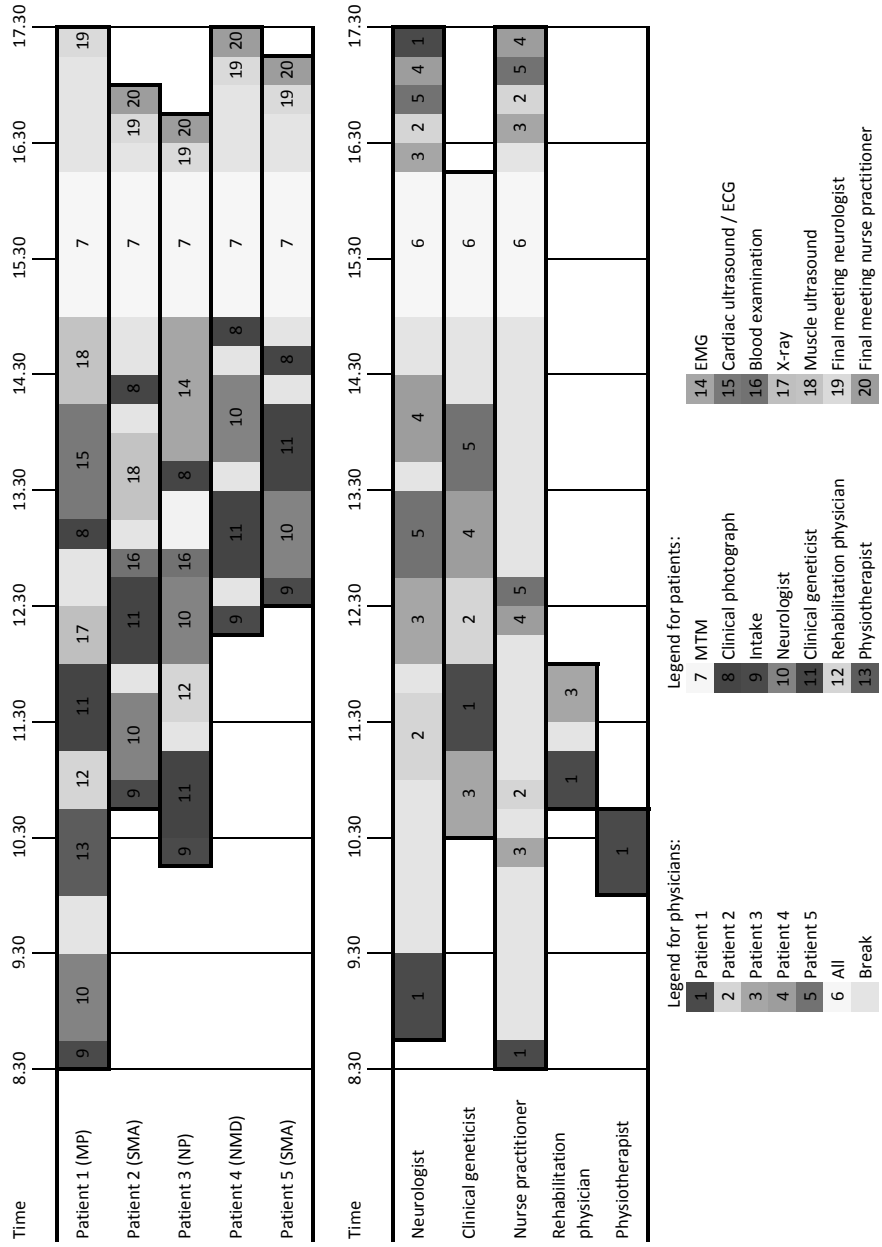


Figure 4.3: An exemplary day schedule for a diagnosis day.

closed during lunch time (12:00–13:00). Therefore, at most two patients can have this examination on one day. When there are five patients on one diagnosis day, just one patient can take the EMG examination, regardless of the other appointments required.

4.4 Access time analysis

Recall that an access time of a patient is defined as the time span from returning the questionnaire until being present at a diagnosis day. In this section, we derive the distribution of access times, assuming that the patients are scheduled in the First-Come First-Served (FCFS) fashion. First, the capacity of a diagnosis day of the CMCA is analyzed. Then, a Markov model is formulated to evaluate access time distribution. Finally, we present numerical results on various demand scenarios.

4.4.1 Number of scheduled patients per diagnosis day

The number of patients that can be scheduled in one diagnosis day is defining for access times. However, due to the complexity of the scheduling problem, this number cannot be directly modeled or predicted, therefore, a simulation study has been performed to determine its distribution. A list of 5900 patients and their arrival times was constructed at random, based on the data given in Table 4.1. Next, for consecutive diagnosis days, the first five patients on the list were selected, and an optimal schedule was constructed. If less than five patients were on the waiting list, then the optimal schedule was constructed for all patients on the list. When a patient was scheduled, he was deleted from the list.

Table 4.4 shows the distribution of the number of patient visits scheduled, given the waiting list size. We schedule the patients in groups of five because, as observed from Table 4.4, there is a high probability that a group of five patients can be scheduled, while it is never possible to schedule six patients. The latter statement can easily be proven by combining the information on the availability of the clinical geneticist, the blood examination, the length of the MTM and the priority of complete visits over partial visits. If two patients are scheduled both have a complete visit. If

Table 4.4: The distribution of the capacity of diagnosis days.

	Number of patients scheduled					
Length waiting list	0	1	2	3	4	5
1	100%	-	-	-	-	-
2	-	-	100%	-	-	-
3	-	-	-	100%	-	-
4	-	-	-	0.3%	99.7%	-
5	-	-	-	0.3%	13.7%	86.0%
6	-	-	-	0.3%	13.7%	86.0%

three or more patients are scheduled, then in almost all cases (99%) three patients have a complete visit and the others have a partial visit. Otherwise, two patients have a complete visit and the others have a partial visit.

As soon as one patient cannot be scheduled while his predecessors have been scheduled, it is obligatory to schedule this patient on the next diagnosis day. Simulations have shown that mainly the needs of four scheduled patients, and not the needs of a fifth patient, determine whether the fifth patient can be scheduled or not. Thus, we can assume that the number of patients scheduled on a diagnosis day depends only on the size of the waiting list, and is independent of how many and which patients were scheduled for other diagnosis days.

4.4.2 Model description

We model the arrivals of new patients as a Poisson process, of which the arrival rate λ (patients per year) is known. The service discipline is FCFS. A year is split into m time periods of equal length, with one diagnosis day per time period. At the opening of the CMCA, the value for m proposed by the AMC is twelve.

Recall the procedure given in Figure 4.1. An access time of a patient consists of three parts. (1) The time until the end of the time period. This time is stochastic and has a uniform distribution. (2) The number of full time periods the patient has to wait until being scheduled. This is stochastic, and has a discrete distribution W which has to be determined. (3) The time between being scheduled and the actual visit to the hospital. This time is deterministic.

Let A_n be the random number of arrivals in time period n. Denote by Q_n the number of waiting patients at the *end* of time period n. Out of Q_n waiting patients, a random number B_n of patients are scheduled for diagnosis day n. The distribution of B_n depends on the value of Q_n. Table 4.4 contains conditional probabilities $P[B_n = b|Q_n = q]$, $q \geq 0$, $0 \leq b \leq q$, for the CMCA case study, obtained from simulations as discussed in Section 4.4.1.

The main performance characteristic of interest is the access time of the patients. Denote by W_n the access time of a patient that arrived at time period n. The distribution of W_n depends on the queue length at the *beginning* of time period n, denoted by Y_n. The following equations hold:

$$Q_n = Y_n + A_n, \tag{4.2}$$

$$Y_{n+1} = Q_n - B_n, \quad n = 1, 2, \ldots. \tag{4.3}$$

Here A_n is independent of the other random variables, and B_n depends on Q_n. Assuming that the arrival rate is not too high, it will often happen that all waiting patients in the queue are scheduled. In that case, the stochastic process W_n will soon reach stationarity, therefore, we choose to obtain its stationary distribution W. To this end, we first determine the stationary distribution Y of Y_n and then obtain the distribution of W using the renewal theory argument. The details of the derivation are provided in Appendix 4.6.2.

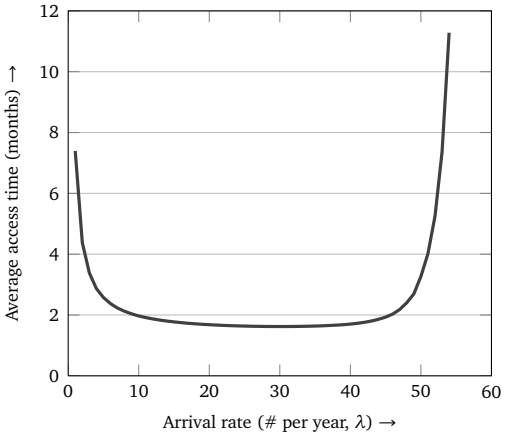


Figure 4.4: Average access times in months (m=12).

4.4.3 Numerical results

The distribution of W is evaluated numerically, by approximating Y_n with a finite Markov chain. This approach is justified by the fact that the queue lengths are typically short. We will present the results computed for our case study. The initial frequency of diagnosis days that the CMCA will apply is (at most) twelve such days per year, therefore, $m = 12$. Next, as stated in Section 4.2, the arrival rate is estimated to be between 20 and 50 patients per year. The time between being scheduled for a diagnosis day and the actual visit to CMCA is precisely one month. For these input parameters we obtain the total average access times, from the time the questionary is received till the hospital visit.

The results are presented in Figures 4.4 and 4.5. From Figure 4.4, we see that excessively large access times are observed in two extreme cases. When the arrival rate is small, less than ten patients per year, large access times arise because at least two patients have to be scheduled on one day, and thus arriving patients often have to wait for another arrival. When there are more than 50 patients a year, large average access times arise since the maximal capacity is almost met. In between, the average access time is stable at a value just below two months. The shape of the distribution of the access times heavily depends on the arrival rate λ. This can be seen in Figure 4.5. When the arrival rate is low, the moment in a month when a patient arrives does not have any influence on the distribution of the access times. However, as the value of λ increases, a heavier dependence shows, because the patients that arrive at the beginning of a month have a considerably higher chance to be scheduled earlier.

We emphasize that the access times are heavily dependent on the constraints of the scheduling problem. When the bottleneck constraints, as mentioned in Section 4.3.2, are relaxed, then large groups of patients can be scheduled together. For example, if the clinical geneticist would be available all day, it will in some cases

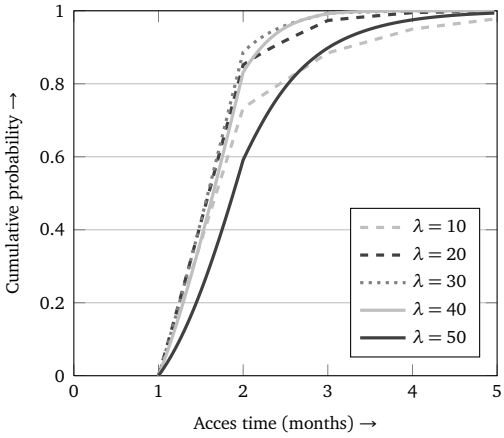


Figure 4.5: Distributions of access times (m=12).

be possible to schedule one more patient per day. This will increase the maximum capacity and result in smaller access times.

4.5 Discussion

We have shown how combining integer linear programming, simulation, and queuing theory helps the AMC in organizing care for children with neuromuscular diseases. The diagnosis and treatment center embodies a transformation from supply-driven to demand-driven patient care. Customized diagnostics and treatment can be offered in a combined visit. To realize this, all practical constraints and preferences were collected and incorporated in an ILP by which feasible day schedules for multiple patient visits can be constructed. Simulations give insight in the capacity of the CMCA, given the availability of staff and equipment and estimates on patient demand in the number of arrivals and required appointments. Finally, a Markov model predicts the access time distributions for diagnostic patients based on the simulation outcomes.

Formulating the day scheduling model was an iterative process, intensively involving the clinicians. Results on initial formulations predicted very long patient access times. Presenting these results to the clinicians, yielded that some highly restricting constraints were loosened or deleted. For example, there was a constraint that the physicians were to have at least two appointments per day. Thus, only patients with the same needs could be scheduled together, resulting in long access times. Another example, where the AMC still struggles with, is the choice whether the MTM should start at a fixed time or not. Although a fixed time is preferred by the clinicians, from a patient's point of view, based on the experimentation outcomes, we strongly recommend it to be flexible. We believe that the benefit of quantitative analysis in such a 'negotiation' process is that it rationalizes the process of realizing

a good trade-off between interests of clinicians and patients.

The main limitation of our study is the availability of accurate data. Since the center just opened its doors, no historical data was available besides data on realizations of how the treatment was previously delivered by different outpatient clinics. Having a focused care center may increase the attractiveness for patients to come to the AMC, which makes predictions on the number of patient arrivals highly uncertain. In addition, patient type mix and required appointments could only be estimated from physician's expert opinions and data on the former patient population. Therefore, we recommend the center to constantly monitor its operations, and to regularly repeat the analysis so to reconsider both the frequency of carrying out treatment days and the staff/equipment availability during a treatment day.

In this study, we have considered the First-Come First-Served discipline for patients to be admitted from the waiting list. There might be a discipline which gives a better performance, if such a discipline increases the number of patients that can be seen on one day. Investigating the existence of such a policy might be an interesting direction for future research. However, when changing the service discipline, the justification for the assumption of independence between the batches of patients scheduled on consecutive treatment days has to be reconsidered. Also, the issue of practical acceptance needs to be addressed, since it questionable how clinically acceptable it would be to not admit the patient that has been on the waiting list the longest.

The first patients visited the CMCA in January 2011 in a pilot phase of the entire treatment concept. During this pilot phase, the nurse practitioner enters the needs of the patients in an Excel sheet. Given a set of patients with prescribed consultations and examinations, and the availabilities of the staff and equipment, the optimal schedule is determined using AIMMS. This is not the desired ultimate state, since it still requires copying the resulting appointments in the electronic agenda system by hand. Also, the AMC strongly opposes the implementation of different software tools in different parts of the hospital, to prevent the maintenance and support task of ICT department to become inefficient if not impossible. Therefore, the scheduling algorithm is intended to be incorporated in the new hospital-wide electronic agenda system that is currently under construction. For such a system, it will be required to be able to communicate with an ILP solver, which will be a main challenge for the ICT design. Modern ICT systems for hospital organization increasingly often embrace OR/MS solutions, in particular, in capacity evaluation and appointment scheduling. In the near future, advanced qualitative schemes, as the one described in this chapter, should become a standard part of hospitals' integral ICT support, for transparent and efficient planning of high quality care.

4.6 Appendix

4.6.1 Mathematical formulation appointment scheduling

This appendix contains the mathematical formulation of the ILP described in Section 4.3.1.

Variables and Parameters

Recall from (4.1) that the decision variables are denoted by $z_{j,p,s,t} \in \{0,1\}$, that equals one if patient j has an appointment with resource s for procedure p starting at time slot t. Besides that, we use several other variables and parameters. As soon as a variable or parameter is used for the first time, it will be introduced briefly. A complete list of sets, indices, variables, parameters and their properties can be found in Tables 4.5 and 4.6.

Constraints

In the ILP, several types of constraints are considered. We distinguish: constraints on the selection of patients, basic planning constraints, precedence constraints, appointment constraints, MTM constraints, and defining constraints. Below each constraint is presented in detail.

Selection of patients. A patients has most of his appointments, all of them, or none. He or she thus has a complete visit, a partial visit, or is not scheduled. We denote by g_j and e_j the binary variables that indicates whether a patient has, respectively, a complete visit ($g_j = 1$) or a partial visit ($e_j = 1$). If $d_j > 1$ then only complete visit is allowed, which will be guaranteed by (4.8). A patient may only have a partial visit if $d_j = 1$. Thus, we only define the variable e_j for patients that have $d_j = 1$. A patient cannot have a partial and complete visit at the same time, therefore we have:

$$g_j + e_j \leq 1 \quad \text{, for all } j.$$

The number of desired appointments that can be skipped is patient specific. This is given by o_j. The binary parameter $N_{j,p}$ denotes whether patient j needs procedure

Table 4.5: Sets and indices ILP.

Set	*Description*	*Index*
J	patients	j
P	procedures	p, p'
S	resources	s
T	time slots	t

Table 4.6: Parameters and variables ILP.

Notation	*Description*
Binary parameters	
$Q_{s,p}$	1 if resource s is qualified to perform procedure p
$A_{s,t}$	1 if resource s is available in time slot t
$C_{p,p'}$	1 if procedures p and p' can be performed simulteneously
$H_{p,p'}$	1 if procedure p has to be performed before procedure p'
$N_{j,p}$	1 if patient j needs to undergo procedure p
$E_{j,p}$	1 if appointment p is necessary for patient j
Integer parameters	
$F_{p,p'}$	minimal number of time slots before start of procedure p' after start of p
t_{MTM}	starting time slot of the MTM
p_{MTM}	procedure number of the MTM
$L_{j,p}$	number of time slots that procedure p takes for patient j. $L_{j,p} = 0$ indicates that procedure p is not required for patient j
m_j	maximum number of appointment time slots patient j can handle on a day
d_j	number of treatment days over which the appointments of patient j may be spread
o_j	maximum number of desired appointments that patient j is allowed to skip in a partial visit
Real parameters	
u_s	relative weight of idle time of staff member s
k_j	relative weight of patient j
Binary variables	
$z_{j,p,s,t}$	1 if patient j has an appointment with resource s for procedure p starting at time slot t
$x_{j,p,t}$	1 if patient j has an appointment for procedure p starting at time slot t
g_j	1 if patient j has a complete visit
e_j	1 if patient j has a partial visit
$b_{j,t}$	1 if patient j has an appointment at time slot t
$c_{j,p,p'}$	1 if patient j has both appointments p and p' scheduled
General integer variables	
y_j^{min}	first time slot at which patient j has an appointment
y_j^{max}	last time slot at which patient j has an appointment
y_s^{min}	first time slot at which staff member s has an appointment
y_s^{max}	last time slot at which staff member s has an appointment

p or not. Denote by the binary $x_{j,p,t}$ whether patient j has an appointment for procedure p starting at time slot t or not. The next constraint only needs to be satisfied if patient j has a partial visit, thus we formulate it as a big-M constraint [637] where

$M_1 = |P|$ satisfies:

$$M_1 \cdot (1 - e_j) + o_j + \sum_{p,t} x_{j,p,t} \geq \sum_{p} N_{j,p} \quad \text{, for all } j.$$

At least two patients should be scheduled on a treatment day, otherwise it is cancelled:

$$\sum_{j} (g_j + e_j) \geq 2.$$

Patients are scheduled according to the FCFS discipline:

$$g_j + e_j \geq g_{j'} + e_{j'} \quad \text{, for all } j,\ j' \in J \text{ such that } j < j'.$$

Basic planning constraints. A patient gets each treatment at most once. Denote by $L_{j,p}$ the number of time slots patient j needs to undergo procedure p. If $L_{j,p} = 0$, the patient does not need the procedure. Thus, we get the following constraint:

$$\sum_{t} x_{j,p,t} \leq 1 \quad \text{, for all } j,\ p \text{ such that } L_{j,p} > 0. \tag{4.4}$$

A treatment is carried out by a resource that has the necessary qualifications. Denote by the binary $Q_{s,p}$ whether resource s is qualified to perform procedure p. Each scheduled procedure should have a qualified resource performing it at the intended time. This constraint, in (4.5), holds for all procedures except the MTM, for which we have a separate set of constraints. Furthermore, (4.5) in combination with (4.4) ensures that no dummy appointments are scheduled to reduce the idle time of staff members:

$$\sum_{s} z_{j,p,s,t} \cdot Q_{s,p} = x_{j,p,t} \quad \text{, for all } i,\ j,\ p \text{ such that } p \neq p_{MTM} \text{ and } L_{j,p} > 0. \tag{4.5}$$

A resource can only be scheduled at one place at the time and only if the resource is available for the CMCA. Denote by the binary $A_{s,t}$ whether resource s is available for the CMCA at time t. Recall that we have defined the starting time of an appointment with $z_{j,p,s,t}$ and the length with $L_{j,p}$, so for each time slot we have to look in the past whether an appointment has started which is not yet finished at the moment:

$$\sum_{j,p} \sum_{t'=t-L_{j,p}+1}^{t} z_{j,p,s,t'} \leq A_{s,t} \quad \text{, for all } s,\ t.$$

Precedence constraints. Some treatments have to be performed before others. For example, the intake appointment of the patient should be before all other appointments of the day. Denote with the binary $H_{p,p'}$ whether procedure p should be

performed before procedure p' in case a patient needs both procedures. Since the constraint only holds for appointments p, p' that are both scheduled, we introduce the binary variable $c_{j,p,p'}$ which is one if both p and p' are scheduled for patient j. This variable is only relevant when both procedures are required by the patient and there is a precedence constraint. The following constraint ensures $c_{j,p,p'} = 1$ when both procedures are required:

$$\sum_t (x_{j,p,t} + x_{j,p',t}) - 1 \leq c_{j,p,p'}, \quad \text{for all } j,\ p,\ p' \text{ such that } H_{p,p'} = 1,\ L_{j,p} > 0,\ L_{j,p'} > 0.$$

Now we can formulate the precedence constraint, in combination with the constraint on a minimum amount of time between the starting times of some combinations of treatments. This minimum amount of time slots is denoted by $F_{p,p'}$. The big-M formulation ensures the constraint is always satisfied when $c_{j,p,p'} = 0$. In this constraint, a value of $M_2 = 3 \cdot |T|$ suffices:

$$\sum_t t \cdot (x_{j,p',t} - x_{j,p,t}) - L_{j,p} - F_{p,p'} + (1 - c_{j,p,p'}) \cdot M_2 \geq 0, \quad \text{for all } j,\ p,\ p' \text{ such that } H_{p,p'} = 1,\ L_{j,p} > 0,\ L_{j,p'} > 0. \tag{4.6}$$

CMCA appointment constraints. A patient can get some of his treatments simultaneously. Denote by the binary $C_{p,p'}$ whether the procedures p and p' can be carried out for the same patient at the same time. The next constraint checks, for each time slot and each combination of appointments, whether they are being performed or not. This is only relevant if the two procedures cannot be performed concurrently, and a qualified resource should be available. This is not relevant for the MTM, since patients are not present there.

$$\sum_{t'=t-L_{j,p}+1}^{t} x_{j,p,t} + \sum_{t'=t-L_{j,p'}+1}^{t} x_{j,p',t} \leq 1, \quad \text{for all } j,\ p,\ p',\ t \text{ such that } C_{p,p'} = 0,\ p > p', \sum_s A_{t,k} \cdot (Q_{p,s} + Q_{p',s}) > 0, p \neq p_{MTM}, p' \neq p_{MTM}.$$

A patient needs a time to rest. Therefore, in the span of three hours, there is at least half an hour free from appointments. These breaks should have the length of at least one quarter of an hour. Denote by the binary variable $b_{j,t}$ whether patient j has an appointment at time t or not. In this variable, the MTM is not considered as an appointment because the patient is not present at the MTM. With time slots of fifteen minutes, this gives the following constraint:

$$\sum_{t'=t}^{t+11} b_{j,t} \leq 10 \quad \text{, for all } j,\ t. \tag{4.7}$$

Note that when time slots have a different length, constraint (4.7) is easily adjusted to ensure the patient has enough time to rest. However, an additional constraint will be neccesary to ensure each break is at least fifteen minutes long.

Some appointments are obligatory: the 'necessary' appointments. If these are not scheduled, the patient cannot visit the CMCA. Denote by the binary $E_{j,p}$ whether an appointment is necessary or not. Thus, for all appointments that are necessary, we require:

$$g_j + e_j \leq \sum_t x_{j,p,t} \quad \text{, for all } j,\ t \text{ such that } E_{j,p} = 1.$$

Patient specific constraints. A patient gets no more treatment time than he can handle on a day. Denote by m_j the maximum number of time slots of appointments that patient j can have on one treatment day. Then we have the following constraint:

$$\sum_t b_{j,t} \leq m_j \quad \text{, for all } j.$$

Sometimes appointments of a patients have to be spread over several days because of the requirements of the patient. The number of treatment days a patient j has left is given by d_j. However, an appropriate fraction of appointment time should be scheduled on the first treatment day in order to avoid scheduling problems on later treatment days. Thus we have:

$$g_j \cdot \sum_{p \neq p_{MTM}} L_{j,p} \leq d_j \sum_t \sum_{p \neq p_{MTM}} x_{j,p,t} \cdot L_{j,p} \quad \text{, for all } j. \tag{4.8}$$

Multidisciplinary Team Meeting constraints. We schedule the MTM as one appointment for all patients. Using precedence constraints below, we will ensure the length of the MTM is correct. The core team members are formally assigned to the first patient:

$$\sum_s z_{j,p,s,t} = x_{j,p,t} \cdot \sum_s Q_{p,s} \quad \text{, for all } t,\ p = p_{MTM},\ j = 1.$$

Now we define the precedence constraints for the MTM. Some of the appointments must be finished before the MTM. Thus, we define a constraint similar to (4.6):

$$\sum_t t \cdot (x_{j',p',t} - x_{j,p,t}) - L_{j,p} - F_{p,p'} + (1 - c_{j,p,p'}) \cdot M_2 \geq 0, \tag{4.9}$$

$$\text{for all } j,\ j',\ p,\ p' \text{ such that } H_{p,p'} = 1,\ L_{j,p} > 0,\ p' = p_{MTM}.$$

Some appointments can start only after the MTM. The length of the MTM for scheduled patient j is $(g_j + e_j) \cdot L_{j,p_{MTM}}$. This yields a constraint similar to (4.9):

$$\sum_t t \cdot (x_{j,p,t} - x_{j',p',t}) - \sum_{j''} L_{j'',p'} \cdot (g_{j''} + e_{j''}) - F_{p,p'} + (1 - \sum_t x_{j',p',t}) \cdot M_2 + (1 - \sum_t x_{j,p,t}) \cdot M_2 \geq 0,$$
$$\text{for all } j,\ j',\ p,\ p' \text{ such that } p' = p_{MTM},\ L_{j,p'} > 0,\ H_{p',p} = 1.$$

Sometimes it is desirable to always start the MTM at a fixed time. Denote by t_{MTM} the time slot in which the MTM should start. Then we obtain a constraint for the starting time of each MTM appointment:

$$\sum_t x_{j,p,t} = t_{MTM} \quad \text{, for all } j,\ p = p_{MTM}.$$

Defining constraints. This group of constraints determines the starting and the end times of the patients and staff members. Denote by y_j^{min} the first time slot when patient j has an appointment. We have to take into account only those appointments that are actually planned, so we construct a big-M constraint. Here $M_3 = |I|$ is sufficient:

$$y_j^{min} \leq M_3 + (t - M_3) \cdot x_{j,p,t} \quad \text{, for all } j,\ p,\ t \text{ such that } p \neq p_{MTM}.$$

The last time slot when patient j has an appointment, y_j^{max}, is determined by the following constraint:

$$y_j^{max} \geq (t + L_{j,p}) \cdot x_{j,p,t} \quad \text{, for all } j,\ p,\ t \text{ such that } p \neq p_{MTM}.$$

Note that when patient j is not planned, y_j^{min} and y_j^{max} can take any integer value in the interval $[0, \ldots, |T|]$.

In a similar fashion we can derive the minimum and maximum values for staff members:

$$y_s^{min} \leq M_3 + (t - M_3) \cdot z_{j,p,s,t} \quad \text{, for all } j,\ p,\ s,\ t;$$
$$y_j^{max} \geq (t + L_{j,p}) \cdot z_{j,p,s,t} \quad \text{, for all } j,\ p,\ s,\ t.$$

The next constraint determines whether a patient has an appointment at a certain time or not, recall that this is denoted by the binary variable $b_{j,t}$. The following constraint forces $b_{j,t} = 1$ when a patient has an appointment. Since a patient can have multiple appointments at one time slot, $M_4 = |P|$ satisfies the following inequality:

$$\sum_{p \neq p_{MTM}} \sum_{t'=t-L_{j,p}+1}^{t} x_{j,p,t'} \leq M_4 \cdot b_{j,t} \quad \text{, for all } j,\ t.$$

The following constraint ensures $b_{j,t} = 0$ whenever patient j has no appointment at time t:

$$b_{j,t} \leq \sum_{p \neq p_{MTM}} \sum_{t'=t-L_{j,p}+1}^{t} x_{j,p,t'} \quad \text{, for all } j,\ t.$$

Objective function

The objective function consists of several parts.

Maximize the number of patients that have a complete visit. Denote by k_j the relative weight of patient j. Then we want to maximize the following expression:

$$\sum_j g_j \cdot k_j. \tag{4.10}$$

Maximize the number of patients that have a partial visit. This expression is similar to (4.10):

$$\sum_j e_j \cdot k_j. \tag{4.11}$$

Maximize the treatment time of all scheduled patients. Note that constraint (4.4) ensures that no dummy appointments are being scheduled. We wish to maximize the total length of all scheduled appointments, so if for example two appointments with lengths $L_{j,p}$ and $L_{j,p'}$ are scheduled at the same time then we need to add $L_{j,p} + L_{j,p'}$ to the total treatment time. Thus, we want to maximize:

$$\sum_{j,p,t} x_{j,p,t} \cdot L_{j,p}. \tag{4.12}$$

Minimize the idle time in the schedules of the clinicians. We have already defined the starting and end times of a staff member. Since the idle time of some staff members (or resources) might be valued different than that of others, we assign a relative weight u_s to the idle time of staff member (or resource) s. Thus, we wish to minimize the following expression:

$$\sum_s u_s \cdot (y_s^{max} - y_s^{min} - \sum_{j,p,t} z_{j,p,s,t} \cdot L_{j,p}). \tag{4.13}$$

Minimize the idle time in the schedules of the patients. It is assumed that this is equally important for all patients. Note that constraint (4.7) ensures that each patient has enough time to rest. Then the total idle time of the patients equals to

$$\sum_j \Big(y_j^{max} - y_j^{min} - \sum_t b_{j,t} \Big). \tag{4.14}$$

The expressions (4.10)–(4.14) contribute to the objective function, each having its own relative importance. The coefficients determining the relative importance are given by α, β, γ, δ and ϵ. Thus, we obtain the following objective function:

$$\max \qquad \alpha \cdot (4.10) + \beta \cdot (4.11) + \gamma \cdot (4.12) - \delta \cdot (4.13) - \epsilon \cdot (4.14). \tag{4.15}$$

4.6.2 Derivation of the access time distribution

In this appendix, we present the derivation of the probability distribution of the stationary waiting time W defined in Section 4.4.2.

First, we write the transition probabilities for Y_n. From (4.3), by conditioning on $[Q_n = q]$ and noting that $P[B_n = q - i | Q_n = q] = 0$ whenever $q - i < 0$, we obtain:

$$P[Y_{n+1} = i | Y_n = j] = \sum_{q=\max(i,j)}^{\infty} P[B_n = q - i | Q_n = q] \cdot P(Q_n = q).$$

Next, using (4.2) we get:

$$P[Y_{n+1} = i | Y_n = j] = \sum_{q=\max(i,j)}^{\infty} P[B_n = q - i | Q_n = q] \cdot P(A_n = q - j).$$

From the transition probabilities above we determine the stationary distribution Y of Y_n. In the case study, we obtain an approximation for the stationary distribution. To this end, we bound the maximal value of Y_n with some large number N so that $P[Y \geq N]$ is sufficiently close to zero. Then the stationary distribution for the bounded chain is computed by numerically solving the balance equations. Finally, we approximate Y with the stationary distribution of the bounded Markov chain, and use $P[Y_n = k] = 0$ when $k \geq N$.

Now our goal is to derive the stationary waiting times. To this end, denote by $P[W \in \mathcal{A} | Y = i]$ the stationary probability that waiting time of an arriving patient is the number in a set $\mathcal{A} \subset \{0, 1, \ldots\}$, provided that there were $i = 0, 1, \ldots$ waiting patients at the beginning of the time slot of the arrival. Consider the sequence of time periods n such that $[Y_n = i]$. The distribution of the waiting times of the A_n patients arriving in such time period is completely defined by Y_n. Thus, given Y_n, these waiting times are independent of the waiting times of the patients arriving in the other time periods. Denote by $A_n^{(k)}$ the number of patients that have arrived in time period n and have to wait k time slots before being scheduled. Then using the renewal reward theory we write:

$$P[W = k | Y = i] = \frac{\mathbb{E}[A_n^{(k)} | Y_n = i]}{\mathbb{E}[A_n | Y = i]} = \frac{\mathbb{E}[A_n^{(k)} | Y_n = i]}{\mathbb{E}[A_n]} = \frac{\mathbb{E}[A_n^{(k)} | Y_n = i]}{\lambda / m}. \quad (4.16)$$

Let us now define the probability $P[W \geq 1 | Y = i]$. From (4.16) it follows that

$$P[W \geq 1 | Y = i] = \frac{\mathbb{E}[A_n - A_n^{(0)} | Y_n = i]}{\lambda / m}, \quad (4.17)$$

where for the numerator we write

$$\mathbb{E}[A_n - A_n^{(0)} | Y_n = i, A_n = k_n] = \sum_{b_n=0}^{i+k_n} P[B_n = b_n | Q_n = i + k_n] \quad (4.18)$$

$$\times \mathbb{E}[A_n - A_n^{(0)} | Y_n = i, A_n = k_n, B_n = b_n], \quad (4.19)$$

and for the last component above holds:

$$\mathbb{E}[A_n - A_n^{(0)} | Y = i, A_n = k_n, B_n = b_n] = \min\{k_n, i + k_n - b_n\}. \tag{4.20}$$

Using (4.17) – (4.20) the probability $P[W \geq 1 | Y = i]$ can be directly computed.

Similarly, we can write the expression for $P[W \geq 2 | Y = i]$. Note that sometimes a patient has to wait longer because there are not enough patients on the list to form a batch of minimal size. Thus, the waiting times of patients arriving in time slot n depend also on the arrivals in time slot $n+1$. Specifically, we derive the following:

$$P[W \geq 2 | Y = i] = \frac{m}{\lambda} \sum_{k_n=0}^{\infty} P[A_n = k_n] \sum_{b_n=0}^{i+k_n} P[B_n = b_1 | Q_n = i + k_n] \\ \times \mathbb{E}[A_n - A_n^{(0)} - A_n^{(1)} | Y = i, A_n = k_n, B_n = b_n],$$

where

$$\mathbb{E}[A_n - A_n^{(0)} - A_n^{(1)} | Y = i, A_n = k_n, B_n = b_n] = \sum_{k_{n+1}=0}^{\infty} P[A_{n+1} = k_{n+1}] \\ \times \sum_{b_{n+1}=0}^{i+k_n+k_{n+1}+b_n} P] B_{n+1} = b_{n+1} | Q_{n+1} = i + k_n + k_{n+1} + b_n] \\ \times \mathbb{E}[A - A^{(0)} - A^{(1)} | Y = i, A_n = k_n, B_n = b_n, A_{n+1} = k_{n+1}, B_{n+1} = b_{n+1}],$$

and

$$\mathbb{E}[A_n - A_n^{(0)} - A_n^{(1)} | Y = i, A_n = k_n, B_n = b_n, A_{n+1} = k_{n+1}, B_{n+1} = b_{n+1}] \\ = \max\{0, \min\{k_n, i + k_n - b_n - b_{n+1}\}\}.$$

In a similar fashion, we derive $P[W \geq k | Y = i]$ for $k = 3, 4, \ldots$. Finally, $P[W = 0 | Y = i] = 1 - P[W \geq 1 | Y = i]$.

We assume that the system functions in a stationary regime, and we use the full probability formula, where the exchange of the limit and the summation is justified by the dominated convergence theorem:

$$P[W \geq k] = \sum_{i=0}^{\infty} P[Y = i] P[W \geq k | Y = i], \quad k = 0, 1, \ldots.$$

This completes the derivation.

Part IV

Coordinating Multidisciplinary Treatments

CHAPTER 5

Scheduling Entire Treatment Plans

5.1 Introduction

Rehabilitation clinics treat patients recovering from injury, illness or disease. Patients require a series of treatments administered by therapists from various disciplines, such as physiotherapy, occupational therapy, social work, psychology, and speech therapy. According to the recent World Health Organization (WHO) report on disability [640], in high-income countries about 18% of the population lives with some form of disability, and the prevalence of disability is rising due to aging populations and the global increase in chronic health conditions. The expenditures for rehabilitation care have substantial pay offs including enhanced economic activity, health outcomes, educational achievements, and participation in community activities of people with disabilities [640]. Public spending on disability programs amounts to 1.2% of GDP for OECD countries and is particularly high in the Netherlands and Norway, where expenditures on disability account for approximately 5% of GDP [640]. The WHO indicates improvement potential of rehabilitation care both in terms of quality and efficiency.

Because rehabilitation care is a multidisciplinary process, coordination within both the care process and the logistical organization is essential [362, 614]. As in many healthcare processes, planning deficiencies have a negative impact on both the quality of care and logistical efficiency [132, 640]. The multidisciplinary nature of the rehabilitation process complicates planning and control. Naturally, the best quality of care is realized when the right treatments are provided at the right time [144]. Rehabilitation care professionals indicate that a short access time [516], a simultaneous start with the various disciplines, and the continuity of the rehabilitation process should be guaranteed. In addition, the complexity of rehabilitation care carries the risk of both undertreatment and overtreatment [522]. Despite the positive cost-effectiveness ratio of current rehabilitation care, both the WHO [640] and a recent improvement program for the Dutch rehabilitation sector [107] observe a large potential for rehabilitation care to be organized more efficiently and effectively. This chapter connects with this improvement potential by presenting a planning methodology that enables the integral planning of multidisciplinary treatment plans. The effectiveness of this planning methodology is demonstrated by its application to a case study in the Academic Medical Center (AMC) Amsterdam.

Considerable enhancements in patient-centeredness, quality of care, and efficiency are achieved. By implementing the methodology, more patients can be treated with the same therapist capacity, and patients benefit from both a higher quality of care and a higher quality of service.

From the WHO report [640], we can conclude that the setting of the rehabilitation clinic at the AMC, and its organizational difficulties and logistical issues at the AMC rehabilitation clinic, are typical of rehabilitation care in general. In current AMC practice, several factors hinder the planning and control of rehabilitation care; of these factors, two main drivers are that planning is decentralized and that computerized support for the planning task is limited. All disciplines, or even therapists, manage their own agendas. Planners are supported by an electronic calendar system. However, the current state of this system comprises a database system that lacks the intelligence of a decision support system (see Section 5.2 for a more detailed discussion). Consequently, in many cases, a short access time and a so-called 'simultaneous start' cannot be realized. Moreover, the timely planning of follow-up appointments can be problematic, which can cause a discontinuity in the rehabilitation process. As a result, certain prescribed treatments may never be realized because they cannot be scheduled. In addition, outpatients have to visit the clinic more often than required, because appointments are spread out over several weekdays instead of combined into a single day. Concerning the system's efficiency, planning deficiencies result in suboptimal use of the valuable time of the therapists. We address these issues by developing a model for planning a series of appointments.

We identify three steps for improving a rehabilitation outpatient clinic's organization. The first step a clinic can take is to obtain insight into the demand and the supply of their rehabilitation care [640]. Although seemingly trivial, this insight is often lacking in practice. A clear perception of demand can be acquired by constructing treatment plans (per disease type or on an individual basis) [148], prescribing all treatments that should be realized during the course of a rehabilitation process. Insight in and control over supply can be gained via centrally managed therapist schedules [483]. As a second step, automated support of the planning task can yield further improvements [77, 640]. A first requirement of a software tool is to enable planners to identify feasible planning proposals for individual patients based on their prescribed treatment plans [107]. Using such a decision support tool, the utilization of therapists could be made clear in an earlier stage, thereby enhancing the planning and control of this precious resource. In a third step, by exploiting operations research techniques, intelligent planning algorithms can be developed and implemented in the software tool to find planning proposals that are efficient for both patients and clinicians. Such tools also permit the evaluation of multiple planning strategies and provide a basis for rationalizing the required number of therapists, aligning therapist agendas, and determining the desired patient mix [369].

This chapter specifically addresses the third step noted above, as we present a methodology for planning series of appointments for rehabilitation outpatients in a multidisciplinary setting, considering the numerous constraints and objectives that apply to rehabilitation treatment planning. By formulating an Integer Linear Pro-

gram (ILP), multiple performance indicators are formulated for planning and are weighted according to a uniform strategy. To incorporate the particular characteristics and preferences of a certain organization, a planning methodology as developed in this chapter needs to be context specific. Our basic approach is generically applicable to the rehabilitation sector, where customization is required for each organization. As we have developed the planning methodology to support the rehabilitation outpatient clinic of the AMC, the ILP was developed in close cooperation with the rehabilitation care experts. The results of the AMC case demonstrate the application of such models for multidisciplinary treatment planning in the rehabilitation sector to be very promising.

Rehabilitation planning has received little attention in the literature. Previous studies have addressed an offline scheduling problem, with a planning horizon of one day or one week for a single discipline [119, 120, 462, 484], or with with considering deterministic demand in a multidisciplinary setting [517]. Planning series of appointments has been addressed for radiotherapy patients in [132, 135] and for chemotherapy patients in [572]. For these patients, treatments must be scheduled during a given number of weeks, strictly taking into account the required rest periods. For a more detailed discussion on the references [119, 120, 132, 135, 462, 484], we refer the reader to the literature review provided in Chapter 4.

This chapter is organized as follows. Section 5.2 describes the case study setting. Section 5.3 presents the ILP model for planning a series of appointments. The planning methodology is applied to data from one of the treatment teams within the rehabilitation outpatient clinic of the AMC. We display the numerical results in Section 5.4, followed by the discussion and conclusion in Section 5.5.

5.2 Background: case study

The rehabilitation outpatient clinic of the AMC employs nine physicians and 30 therapists of various disciplines, who jointly perform approximately 10.000 consultations a year. Since 2008, the clinic has participated in an improvement program for the administration and planning practice by implementing a complete package of process redesign interventions, of which we will mention the main two. First, the agenda management was centralized, and uniform schedules for the therapists were created. Second, standard treatment plans were formulated to standardize care processes, to prevent undertreatment and overtreatment, and to obtain insight into demand. These two interventions are the starting point for the work presented in this chapter, which introduces a planning methodology to enable optimal scheduling of the series of appointments prescribed in a treatment plan.

The patient flow, which is changing due to the current introduction of treatment plans, is displayed in Figure 5.1. In the situation of 2008, the rehabilitation process started with a so-called intake consultation with a rehabilitation physician, who decided on the disciplines that should be involved in the patient's care. The therapists determined the frequency and the timing of the treatments. After several weeks, the rehabilitation physician and the therapists discussed the condition of

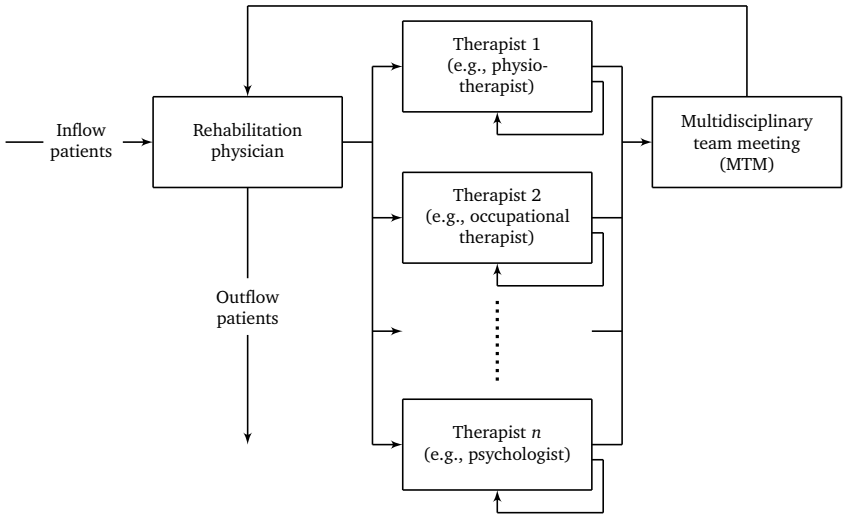


Figure 5.1: Patient flow diagram.

the patient during a Multidisciplinary Team Meeting (MTM). Together, they either decided to terminate or to continue the treatment. As therapists strive to provide patients with the best possible care, the clinicians did report a risk of overtreatment. For each discipline, a follow-up appointment for the patient was only scheduled after the current treatment had taken place, resulting in scheduling on short notice. As this policy hampers the scheduling of an appointment at the prescribed moment, appointments were often scheduled later than prescribed, whereas the scheduling of certain appointments was omitted, thus resulting in undertreatment.

The introduction of treatment plans is changing the patient flow. Following the intake consultation, the rehabilitation physician designs a treatment plan. The standard treatment plans form the basis for each patient treatment. In addition, physicians have the freedom to customize treatment plans if induced by individual patient needs. The treatment plan prescribes the disciplines that should be involved in the patient's treatment, the required number of treatments per discipline, the duration of each treatment and the week in which it should take place. Subsequently, all treatments up until the first MTM are scheduled according to the treatment plan. During the MTM, the rehabilitation physician and the therapists decide either to terminate the treatment of the patient or to design a plan for the continuation of the treatment. In the latter case, the required treatments are scheduled and the patient is scheduled to be discussed again during one of the upcoming MTMs.

Since January 2009, therapists and physicians of the rehabilitation outpatient clinic are grouped in three diagnosis-related treatment teams: Team Paediatrics, Team Neurology, and Team Orthopedics & Traumatology. Each team has a dedicated planner who manages the schedules of all team members, so that treatment planning is centralized. Therapist schedules are standardized such that the time for

patient care and the time for meetings or administration are synchronized among all therapists insofar as possible. Planners use the electronic calendar system X/Care (McKesson) to register appointments and select free appointment slots; therefore, planning is partially automated. However, X/Care has no flexible functionality for scheduling treatment plans, let alone generating efficient planning proposals. When planning a treatment plan, planners have to consider the availability of therapists and of the patient in addition to patient preferences. Hence, whereas a single feasible planning proposal is already difficult to find, the planning task is further complicated by a complex set of constraints and preferences (see Section 5.3). Thus, finding a planning proposal for a complete treatment plan is a very time-consuming and cumbersome task. Planners indicate that they spend on average 15 minutes to find one feasible planning proposal for a multidisciplinary series of treatments for a patient. Therefore planning requests cannot be dealt with immediately. Instead, planners tend to save up and execute planning requests once a week.

When the planner finds a feasible planning proposal, the appointments are fixed and the patient is informed via a letter. This process leaves very little room for patient preferences and is therefore not patient-centered. Moreover, if the patient is not available at some of the appointment times, the patient has to call the rehabilitation outpatient clinic and the planner has to reconsider the planning request. Some patients simply do not show up for their appointments without calling to cancel; it may be that such patients have not received the letter. The ability to execute a planning request promptly, when the patient is on the phone or at the desk, would leave more room to incorporate patient preferences, result in time savings for planners, and presumably reduce the number of no-shows.

In September and October 2009 we performed baseline measurements of two performance indicators for all new patients starting their rehabilitation process (70 patients). As not all required information was available from the hospital databases, the rehabilitation planners manually registered the access time of each new patient and we assessed the case history of each individual patient. The average utilization of therapists during this period was 69%, and the average utilization per discipline differed considerably (see Table 5.5). An access time within two weeks was achieved only for 22.9% of the patients. Of the 38 patients who required treatment with more than one discipline, 52.6% had a simultaneous start with the various disciplines. (For the exact definition of these performance indicators, see Section 5.3.1).

Given the observations described, the current problems described in Section 5.1, and the results of the baseline measurements, it is to be expected that an intelligent planning methodology providing online decision support for the planners would be highly valuable to the rehabilitation outpatient clinic of the AMC.

5.3 Methods

In this section, the planning methodology is presented. First, the requirements of the model and the performance indicators are described, followed by the model formulation. The detailed mathematical formulation of the model is displayed in

the appendix. Here, we discuss the framework of the model by describing the decision variables, the constraints, and the objective function. Figure 5.2 displays an overview of the model.

5.3.1 Requirements of the model

Given a patient with a prescribed treatment plan, and the skills and availabilities of the therapists, the model has to generate a planning proposal consisting of an assigned therapist and a start time for each appointment. The planning proposal, which must comply with the restrictions and preferences of the rehabilitation department, should result in a high-quality schedule for both the patient and the therapists involved.

In close cooperation with the clinicians of the rehabilitation outpatient clinic, we have formulated five performance indicators for the planning methodology, which are defined as follows:

- *Access time.* The number of days from the registration of a patient until the first appointment.
- *Simultaneous start.* The first appointments of a patient with the various disciplines take place within a pre-specified period (e.g., five working days).
- *Lead time.* The number of days from the first until the last appointment of a patient.
- *Combination appointments.* The number of days a patient has to visit the outpatient clinic compared to the minimal number of days necessary.
- *Therapist utilization.* The percentage of time available for patient care that is actually utilized for appointments.

In certain cases, a series of appointments can only be scheduled if some prescribed appointments are omitted. Because rejecting a planning request is far less desirable than omitting a small number of appointments, we allow for these appointments to not be scheduled if their number does not exceed a certain ratio per discipline (see Appendix 5.6). Moreover, clinicians indicate that quality of care cannot be guaranteed when the access time exceeds a certain threshold. To guarantee quality of care, a patient is referred to another clinic if the access time exceeds this threshold (see Appendix 5.6). Of course, it is highly preferable to reduce both of these occurrences to a minimum. Therefore, we also evaluate the performance of the following two indicators:

- *Referred patients.* The percentage of patients referred to another clinic.
- *Unscheduled appointments.* The percentage of appointments prescribed but not scheduled.

5.3.2 Model formulation

To obtain an optimization problem of manageable dimensions for which a provably optimal solution can be found within reasonable time, we model the rehabilitation treatment planning problem as ILP. In an ILP, restrictions specific to the rehabilitation treatment planning problem can be modeled appropriately, and multiple objectives can be weighted rationally.

The ILP is intended for scheduling a series of appointments for one patient at a time. Although this process may not produce the best overall schedules, it enables a direct response to a patient issuing a planning request, which is strongly preferred by the AMC for patient-centeredness reasons. For each series of appointments, the treatment plan prescribes the required number of treatments per discipline, the duration of each treatment, and the week in which it should take place. For each discipline, all appointments should be with the same therapist to ensure continuity of care. Scheduling a series of appointments exactly as prescribed by the treatment plan may not always be possible. Because rejecting a planning request is far less desirable than scheduling a series of appointments in a way that slightly deviates from the treatment plan, we allow for some scheduling flexibility. First, if an appointment cannot be scheduled in the week(s) prescribed by the treatment plan, it may be scheduled a week earlier or later if these weeks do not already contain appointments with the same discipline. Second, as pointed out in Section 5.3.1, if the series can be scheduled except for a few appointments, we allow these appointments to not be scheduled if their number does not exceed a certain ratio per discipline. If a series cannot be scheduled despite this flexibility, we shift the planning horizon one week ahead and try again to schedule the series of appointments.

After each series of appointments, the patient is discussed during an MTM, in which the decision is made either to terminate or to continue the treatment. In the latter case, another series of appointments needs to be scheduled after the MTM. When scheduling the next series, information about the previous series may be relevant. This situation is described in detail in the appendix.

Decision variables. For each appointment within a series, we have to decide on the assigned therapist and the starting time slot. We use the index a for appointments, h for therapists, and t for time slots. The decision variables are as follows:

$$x_{aht} = \begin{cases} 1 & \text{, if appointment } a \text{ is assigned to therapist } h \\ & \text{and starts in time slot } t, \\ 0 & \text{, otherwise.} \end{cases}$$

Constraints. We distinguish several types of constraints:

Basic planning constraints. Appointments may not overlap, both the therapist and the patient have to be available for an appointment, and precedence relations between appointments must be satisfied.

Unscheduled appointments. For each discipline, a maximum of one in every R appointments may be left unscheduled.

Therapist assignment. Per discipline, all appointments must be assigned to the same therapist. This so-called longitudinal continuity of care is a means of improving patient satisfaction and the outcomes of care [34].

Number of appointments per period. Multiple appointments with the same therapist may not be scheduled on the same day. Preferably, multiple appointments with one therapist are spread out evenly, both within and over weeks. The number of appointments with one therapist in a week is limited to L, and the number that may be scheduled on a single day is limited to K.

Start of the rehabilitation process. The access time of the patient should preferably be within S weeks and may not exceed $C \cdot S$ weeks. To realize a simultaneous start, it is preferable that the first appointment with each discipline takes place within V days of the patient's very first appointment.

Continuity of the rehabilitation process. An appointment should preferably be scheduled in the range of weeks prescribed by the treatment plan. However, it may be scheduled a week earlier or later if these weeks do not already contain appointments with the same discipline.

Patient preferences. Because combination appointments are high on the list of outpatient preferences [603], we strive to schedule the appointments on as few days as possible. The waiting time between appointments on the same day may not exceed U time slots.

Recurring day and time. It is preferable that the appointments take place on the same day and time each week such that the patient has fewer days and times to remember.

Efficient filling of therapist schedules. We aim to schedule appointments right at the start or at the end of a session of the therapist, or right before or after an already scheduled appointment. This process prevents a break in the schedule between two consecutive appointments, that might be too short to fit in another appointment. Hence, we thereby minimize the number of referred patients and unscheduled appointments.

Objectives. The objective of the ILP is to minimize the sum of weighted penalty costs. Each of the penalized situations is described below and is characterized by one or more specific constraints in the appendix, referenced by the numbers in Figure 5.2. The objective function consists of two main components. First, it contains terms for the identified performance indicators (see Section 5.3.1), and, second, terms for three additional undesired situations. Let us start with describing the objectives that correspond to the performance indicators:

- *Access time.* The number of time slots by which the preferred access time is exceeded (weight factor α)
- *Simultaneous start.* No simultaneous start realized with the various disciplines (weight factor β)

min	$\eta \cdot$	# unscheduled appointments	(5.4)
	$+\kappa \cdot$	# appointments not spread evenly over the week (per discipline)	(5.9)
	$+\alpha \cdot$	exceeding the preferred access time	(5.12),(5.14)
	$+\beta \cdot$	not starting simultaneously	(5.13),(5.16)
	$+\theta \cdot$	deviation of appointments from the week(s) prescribed in the treatment plan	(5.21),(5.22)
	$+\gamma_1 \cdot$	scheduled duration exceeds prescribed duration by 2 weeks or less	(5.23)
	$+\gamma_2 \cdot$	scheduled duration exceeds prescribed duration by between 1 and 2 weeks	(5.23)
	$+\gamma_3 \cdot$	scheduled duration exceeds prescribed duration by more than 2 weeks	(5.23)
	$+\delta \cdot$	extra appointment days (i.e., rather than combination appointments)	(5.24)
	$+\chi \cdot$	# non-recurring starting time slots	(5.26),(5.27)
	$+\zeta \cdot$	# appointments causing break in the schedule of a therapist	(5.28)–(5.30)
s.t.	no overlapping appointments		(5.1)
	therapist and patient available during the appointment		(5.2)
	precedence relations		(5.3)
	at most 1 out of R appointments per discipline unscheduled		(5.5)
	appointments per discipline always with same therapist		(5.6),(5.7)
	at most 1 appointment per therapist per day		(5.8)
	at most L appointments with 1 therapist in a week		(5.10)
	at most K appointments per day		(5.11)
	exceeding of access time $\leq$ maximum allowed exceeding		(5.15)
	appointments at most one week earlier or later than prescribed		(5.17)–(5.20)
	time between consecutive appointments in one day $\leq U$		(5.25)

Figure 5.2: Overview of the ILP (the numbers refer to the corresponding constraints in Appendix 5.6).

- *Lead time.* The number of weeks by which the prescribed total duration of the series of appointments is exceeded (weight factor γ_1 in case of exceeding by two weeks or less, γ_2 in case of exceeding between one and two weeks, and γ_3 in case of exceeding by more than two weeks)
- *Combination appointments.* The number of extra days the patient has to visit the outpatient clinic because combination appointments have not been scheduled optimally (weight factor δ)
- *Therapist utilization.* The number of breaks created in the therapists' schedules (weight factor ζ)
- *Unscheduled appointments.* The number of unscheduled appointments (weight factor η)

The performance indicator *referred patients* is not contained in the objective function, because patients is only referred when there are no feasible solutions. In addition to penalizing situations not adhering to the performance indicators, we penalize for three additional (undesirable) situations:

- *Compliance to treatment plan.* The number of appointments that are scheduled a week earlier or later than prescribed in the treatment plan (weight factor θ)

- *Appointment spread.* The number of appointments that take place one day after a previous appointment with the same therapist, such that the appointments per discipline are not spread out evenly over the week (weight factor κ)
- *Recurring appointment time.* The number of unique (i.e., non-recurring) appointment times (weight factor χ)

One may observe that the objective function contains multiple goals that are possibly in conflict. For example, in some cases, it is possible to either schedule the first appointment within the preferred access time or to provide the patient a simultaneous start, but not both. As a second example, to optimally schedule combination appointments, it may be beneficial not to schedule certain appointments. By varying the weight factors, the relative importance of the various goals can be specified. The values of the weight factors can be set according to the preferences of the rehabilitation clinic in question. For each clinic, setting these values is part of configuring the ILP to the specific situation.

5.4 Numerical results

5.4.1 Description of the test cases

In this section, we apply the planning methodology to Team Neurology of the rehabilitation outpatient clinic in the AMC. Team Neurology mainly treats patients suffering from neuromuscular diseases, amyotrophic lateral sclerosis, post-polio syndrome, and cerebrovascular accidents.

After the intake consultation, the rehabilitation physician can assign the patient to a treatment plan in two ways. First, he can design an individual treatment plan for each new patient. Second, treatment plan blueprints were formulated by the clinicians of the rehabilitation clinic such that he can assign each new patient to one of the blueprints. We test the methodology with seven treatment plan blueprints formulated by rehabilitation professionals. Table 5.1 shows the characteristics of these seven treatment plan blueprints. Each patient in our experiments is assigned to one of these seven blueprints. The relative frequency of the blueprints is based on hospital database information.

As Team Orthopedics & Traumatology employs no psychologist, patients from Team Orthopedics & Traumatology needing psychology are treated by the psychologist of Team Neurology. To represent the influence of care demands from these patients, we introduce a dummy treatment plan (see Table 5.1). As we do not incorporate the entire treatment plan of these patients because they are not assigned to Team Neurology, we do not include them in the summary scores on the various performance indicators.

Team Neurology employs nine therapists. Table 5.2 displays the availability of each therapist for direct outpatient care. Therapists spend their remaining time on indirect outpatient care (e.g., writing reports and ordering rehabilitation aids), meetings, inpatient care, and research. Since time for these activities is specifically

Table 5.1: Characteristics of the treatment plan blueprints.

Treatment plan	*Patients*	*Series*	*Required*	*Duration*	*# Appointments per discipline (# hours)*				
					PT	*OT*	*ST*	*SW*	*PS*
Amyotrophic lateral sclerosis	22%	1	100%	5	3 (3.0)	4 (3.5)	3 (3.0)	1 (1.0)	
		2	40%	8	4 (3.5)	2 (2.5)	5 (5.0)	1 (1.0)	
		3	20%	5	2 (1.5)	1 (1.5)	1 (2.0)	4 (3.5)	
Post-polio syndrome	13%	1	100%	2	3 (2.5)	1 (1.0)			
		2	60%	2	2 (1.5)	1 (1.0)			
		3	20%	3	1 (2.0)	4 (5.5)			
Neuromuscular diseases (other)	4%	1	100%	4	4 (4.0)	1 (1.5)		1 (1.0)	
		2	50%	6	1 (1.5)	1 (2.5)		2 (1.5)	
		3	20%	10	2 (1.5)	3 (3.0)		2 (1.5)	
Cerebrovascular accidents	17%	1	100%	3	3 (3.0)	2 (3.0)		1 (1.0)	1 (1.0)
		2	50%	7	4 (2.0)	2 (3.0)		2 (3.0)	3 (2.5)
Physiotherapy only	16%	1	100%	2	2 (1.5)				
		2	70%	2	1 (0.5)				
		3	50%	4	2 (1.5)				
		4	30%	5	2 (1.5)				
Occupational therapy only	23%	1	100%	1		1 (1.0)			
		2	50%	4		2 (3.5)			
		3	25%	4		2 (3.0)			
Ortho-trauma dummy	5%	1	100%	4					4 (4.0)

Explanation of the column items

Treatment plan:	name of the treatment plan
Patients:	percentage of patients assigned to this treatment plan
Series:	number of the series of appointments within a treatment plan
Required:	after each series of appointments, during an MTM the decision is made either to continue or to terminate the treatment of the patient; displayed is the percentage of patients continuing for the indicated series
Duration:	prescribed duration in weeks of the series of appointments
# Appointments per discipline:	number of appointments within the series, for each discipline, including the total duration
PT	physiotherapy
OT	occupational therapy
ST	speech therapy
SW	social work
PS	psychology

reserved in their agendas, the sessions during which a therapist is indicated to be available for direct outpatient care are preferably completely filled with appointments. In Table 5.2, morning sessions last from 9:30 until 12:30 and afternoon sessions from 13:30 until 16:00. Therapists are not necessarily available for a full session. An indicator of therapist availability in Table 5.2 means that the therapist is available for at least one hour during that session. As therapists are not always

Table 5.2: Weekly agenda of the therapists' availability for direct outpatient care.

	Monday		*Tuesday*		*Wednesday*		*Thursday*		*Friday*		
Therapist	a.m.	p.m.	a.m.	p.m.	a.m.	p.m.	a.m.	p.m.	a.m.	p.m.	*# Hours*
Physiotherapist 1											18
Physiotherapist 2											17
Occupational therapist 1											13
Occupational therapist 2											6
Occupational therapist 3											13
Occupational therapist 4											6
Speech therapist											14
Social worker											14
Psychologist											10

available for outpatient care, certain (combination) appointments can only be made on specific days or at specific moments, which is quite restrictive for planning.

Table 5.3 lists the values used for the parameters in our experiments, which we set according to the restrictions and preferences of the AMC rehabilitation outpatient clinic. To be able to evaluate performance of the planning methodology from an organizational point of view, in our experiments we assume that patients are always available. All appointments have a duration that is a multiple of 30 minutes. Therefore, in the experiments, each time slot has a length of 30 minutes.

Table 5.4 lists the values used for the weight factors in the experiments. To determine these values, the clinicians of the rehabilitation outpatient clinic scored the relative importance of each part of the objective function. As certain variables are binary whereas others are integer, we applied a normalization factor to each variable in order to generate comparable measures. These normalization factors, multiplied by their relative importance, produced the weight factor values listed in Table 5.4.

5.4.2 Experimental setup

We use discrete-event simulation to evaluate the performance of the presented planning methodology. Prior to the actual simulation, we generate patient arrivals according to a Poisson process. The arrival rate of the Poisson process is set such that a desired therapist load is generated. For each patient, the release date and all treatment requirements are stored in a database. These requirements are generated based on the percentages listed in Table 5.1. Each patient is randomly assigned to one of the seven treatment plan blueprints. In addition, the required number of appointment series is drawn.

During the simulation, the patient with the earliest release date is selected from the database, and appointments are scheduled for this patient. Subsequently, the

performance indicators are updated, the release date of the patient is set to the date of the MTM in which the patient will be discussed, and the next patient is selected. As patients entering the system near the end of a simulation run cannot finish their treatment before the end of the run, we exclude the results of patients arriving during the last 20 weeks, which is the duration of the longest treatment plan.

We evaluate three scenarios. First, the base scenario, with an average therapist load of 70%, is comparable to the therapist load during the baseline measurement observation period. To investigate the potential of the planning methodology to facilitate growth in demand, the average therapist load is set to 80% and 90% for the second and third scenarios, respectively. The average *therapist utilization* may differ slightly from the average therapist load due to three factors: first, the variation in the generation of patient arrivals; second, the percentage of *unscheduled appointments*; and third, the percentage of *referred patients*, with the latter two being preferably minimal.

Based on an analysis of the first five performance indicators (see Section 5.3.1) for five test runs, we set the warm-up period and the run length. The warm-up period is determined by applying Welch's procedure [383] and is set to 2 years. This relatively long warm-up period results from the fact that the simulation starts from an empty system, whereas treatment plans have an average duration of 6.2 weeks, with the longest plan being 20 weeks. The run length (including the warm-up period) is set to 12 years. Based on a desired half-width of 5% for the 95% confidence intervals of the performance indicators *simultaneous start*, *lead time*, *combination appointments*, and *therapist utilization* and a desired half-width of 10% for the 95% confidence interval of the performance indicator *access time*, the number of replications is set at 7 for Scenarios 1 and 2, and at 10 for Scenario 3.

The ILP was implemented in ILOG OPL 6.3 and solved using CPLEX 12.1. For our experiments we used a 2.27 GHz Intel Core i3 ASUS Notebook with 4 GB RAM under a 64-bit version of Windows 7. Because the ILP is intended for scheduling a series of appointments for one patient at a time, numerous ILP instances must be

Table 5.3: Parameter values.

Parameter	*Description*	*Value*
D	number of time slots per day	13
R	number of appointments per discipline, of which at most one may be unscheduled	5
L	maximum allowed number of appointments with one therapist in a week	3
K	maximum allowed number of appointments on a single day	3
S	number of weeks of preferred maximal access time	2
W	number of time slots per week	65
C	factor by which the exceeding of the access time is limited	1
V	number of days within which all first appointments preferably take place (simultaneous start)	5
T	number of time slots in the planning horizon	325
U	maximum allowed waiting time between two consecutive (combination) appointments on a day	1

Table 5.4: Weight factor values.

Weight factor	*Objective*	*Value*
α	access time	20
β	simultaneous start	200
γ_1	lead time	50
γ_2	lead time	150
γ_3	lead time	300
δ	combination appointments	20
ζ	therapist breaks	5
η	unscheduled appointments	500
θ	deviation from treatment plan	1
κ	spreading of appointments	1
χ	recurring day and time	0

solved during a simulation run. Most instances are solved to optimality within a few seconds. The average solving time is 14.2 seconds in Scenario 1 and decreases with increasing load, resulting in an average of 3.1 seconds for Scenario 3. In exceptional cases it can take several minutes to solve to optimality. This prolongation occurs in some of the cases in which a new multidisciplinary patient issues a planning request but therapist utilization is relatively low. Because the therapists to whom a new patient will be assigned have to be decided on and the therapist utilization is relatively low, the solution space is large in such cases.

To control the total duration of a simulation run, a CPU time limit of 600 seconds is applied to each ILP instance. Less than 0.005% of all instances are actually affected by this time limit. Hence, an optimal solution is identified in almost all cases, and for the remaining instances a near optimal solution is generated.

5.4.3 Results

Table 5.5 shows the experimental results for the three scenarios compared to the results of the baseline measurements. Clinicians are highly satisfied with the planning proposals generated by the model. The proposals generated are immediately implementable, without adjustment.

The planning methodology developed relates to the modified patient flow entailed by the introduction of the treatment plans (see Section 5.2). For the rehabilitation outpatient clinic, this new system differs so substantially from current practice, that there is no point in comparing the planning proposals generated by the model with the schedules that are currently being produced by the planners manually. Hence, the best we can do is to compare the results for the performance indicators realized by the model with the baseline measurements.

Note that the objective function of the ILP is the mechanism to direct the scheduling of appointments per individual patient. The value of the objective function in itself is insignificant because we are interested in the realized planning product for the total patient population, which is evaluated by means of the formulated performance indicators. Results for the performance indicators *simultaneous start* and

Table 5.5: Results of planning methodology compared to current practice.

Performance indicators	*Baseline measurements*	*Scenario 1 (load 70%)*	*Scenario 2 (load 80%)*	*Scenario 3 (load 90%)*
Access time *% of patients with an access time ≤ 2 weeks*	22.9%	98.9%	89.5%	53.7%
Simultaneous start *% of multidisciplinary patients having a simultaneous start*	52.6%	100.0%	98.2%	90.8%
Lead time *% of patients with a lead time ≤ 10% longer than the prescribed duration*	n.a.	92.6%	84.1%	69.3%
Combination appointments *% of combination appointments offered to multidisciplinary patients*	n.a.	99.1%	97.4%	93.4%
Therapist utilization - overall *% of time available for patient care utilized for appointments*	69.3%	70.1%	79.3%	87.4%
Per discipline: PT	72.3%	73.1%	83.2%	92.2%
OT	72.1%	73.0%	83.0%	91.1%
ST	74.5%	75.0%	82.4%	88.9%
SW	60.7%	61.6%	69.7%	77.5%
PS	53.3%	53.6%	61.5%	68.9%
Referred patients *% of patients referred to another clinic*	n.a.	0.00%	0.29%	2.47%
Unscheduled appointments *% of appointments not scheduled*	n.a.	0.12%	0.25%	0.33%

combination appointments only apply to patients being treated by multiple disciplines, and are therefore only reported for these patients. As can be observed from Table 5.1, 56% of all patients follow a multidisciplinary treatment plan.

For four of the performance indicators, the results of the baseline measurements are not available for various reasons. During the baseline measurement observation period, the preferred duration of the rehabilitation process of a patient was not prescribed, such that we had no benchmark for the *lead time*. As appointments were scheduled one by one, it was hard to reconstruct which appointments could have been scheduled on the same day, complicating the measurement of the percentage of *combination appointments*. Because *referred patients* and *unscheduled appointments* were also not registered under the old system, these indicators were also unable to be measured during the baseline period.

The results of the baseline measurements and the experiments are displayed in Table 5.5 and Figures 5.3 and 5.4. With a *therapist utilization* comparable to the baseline measurements, the percentage of patients with an *access time* within two weeks increases from 22.9% to 98.9%, representing an improvement of 76%. The percentage of patients with a *simultaneous start* also improves from 52.6% to 100.0%. Additionally, in nearly all cases (99.1%), *combination appointments* are

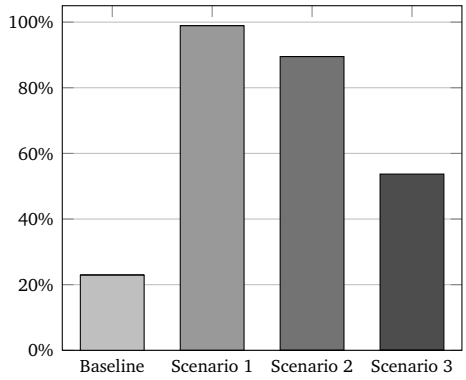


Figure 5.3: Percentage of patients with an *access time* within 2 weeks

offered to patients. Although the results for *lead time* cannot be compared to the baseline measurements, based on the experiences of our clinicians we can state that the results of the experiments significantly outperform current practice; in addition, undertreatment is prevented. As strongly preferred, the percentages of *referred patients* and *unscheduled appointments* are very low.

When the therapist load is increased, the methodology still results in the production of a high-quality plan. With a therapist load of 80%, *simultaneous start* and *combination appointments* have values above 95%, and *access time* and *lead time* have values of 89.5% and 84.1%, respectively. With a further increased therapist load of 90%, *simultaneous start* and *combination appointments* continue to perform very well. However, *access time*, *lead time* and *referred patients* begin to deteriorate. To address this degradation in performance, we suggest three possible actions. First, a simple intervention to improve the continuity of care would be to discuss the patient during an MTM in the week before the last scheduled appointments. In that way, the scheduling of follow-up appointments, if necessary, can take place a week earlier. Second, the values for weight factors in the objective function of the ILP might be adjusted, presumably at the cost of the other performance indicators. As pointed out earlier, in the end it is up to the healthcare professionals to decide on the relative importance of the different performance indicators. Third, by reserving future capacity for patients already under treatment and requiring follow-up appointments, or for new patients, *access time*, *lead time*, and *referred patients* can possibly be improved. However, developing good reservation schemes is a study in itself, as the effects of reserving capacity on the various performance indicators are not trivial. Notably, with a *therapist utilization* of 87.4%, the model in its current form significantly outperforms the baseline measurements, which are realized at a *therapist utilization* of 69%. Hence, by implementing the planning methodology, more patients can be treated with the same therapist capacity, and patients are offered both a higher quality of care and a higher quality of service.

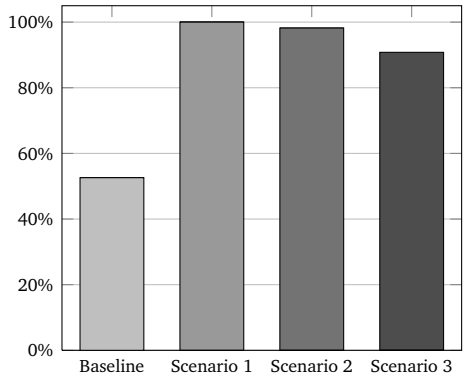


Figure 5.4: Percentage of multidisciplinary patients with a *simultaneous start*.

5.5 Discussion

In this chapter, we have presented a methodology for scheduling series of appointments for rehabilitation outpatients, that improves both the quality of care and logistical efficiency. These improvements in quality of care are realized through significantly shorter access times, an increased percentage of simultaneous starts, an enhanced continuity of care, a better coordination between disciplines via the introduction of treatment plans, and the elimination of undertreatment and overtreatment. These findings are supported by the numerical results of a case study within the rehabilitation outpatient clinic of the AMC.

The planning methodology enhances patient-centeredness as it improves quality of care, provides patients with quick service, and yields a high percentage of combination appointments. Moreover, patient preferences, such as longitudinal continuity of care, are incorporated in the model. Multiple planning proposals can be generated quickly so that the patient is presented with a number of proposals to choose from. Different planning proposals can be generated by varying patient availability or by varying the weight factor values. Because a planning proposal can be generated within seconds, the model can deal with a planning request online, whereas, currently, planners tend to save up planning requests and execute the time-consuming and cumbersome planning task once a week. Dealing with a planning request on the fly reduces access times and provides prompt service to patients and up-to-date insight in terms of the demand for the rehabilitation clinic. This approach also presumably reduces the number of no-shows because patients are unquestionably notified of their appointments, and patients can immediately verify whether or not they are available at the proposed appointment times. Furthermore, the methodology induces cost savings as it reduces the time rehabilitation planners spend per planning request. Planners spend on average 15 minutes to put together one feasible planning proposal for a multidisciplinary series of treatments for a patient, whereas the model generates such a proposal within seconds.

Current healthcare planning systems do not support integral treatment planning. We have developed a prototype of a tool that does support such planning, and we have tested it in a rehabilitation outpatient clinic. Both patients and clinicians are highly satisfied with the planning proposals generated by the model. This would not have been possible without formulating the model in cooperation with physicians, therapists, planners, and management of the rehabilitation outpatient clinic. Thus, despite the wide range of objectives and constraints, by carefully investigating these and formulating these in an ILP, our study has demonstrated that automated support of the planning task is possible. Based on the workability and the expected performance, the management of the AMC has decided to include our planning methodology in the new hospital information system.

Planning multidisciplinary treatments is complex. The multidisciplinary character of rehabilitation care entails interaction between the agendas of the various therapists. The treatment of a patient with a particular discipline can only begin once the other disciplines required also have available capacity, and during the rehabilitation process appointments with the various disciplines have to be synchronized. As this interaction influences all performance indicators, aligning the capacities of the disciplines is of utmost importance. For the AMC case, the imbalance between the utilizations per discipline (see Table 5.5) may have a negative impact on the results, especially when therapist load is high, as an overloaded discipline blocks multidisciplinary patients from entering the clinic, whereas at the same time the other disciplines might have capacity available to accept those patients.

The AMC case is relatively small, with three disciplines (speech therapy, social work, and psychology) consisting of only one therapist. Although a larger case presumably results in a longer computation time, it increases planning flexibility, likely resulting in improved schedules. For example, there would be more freedom to select the therapists to whom the patient could be assigned, and as each discipline

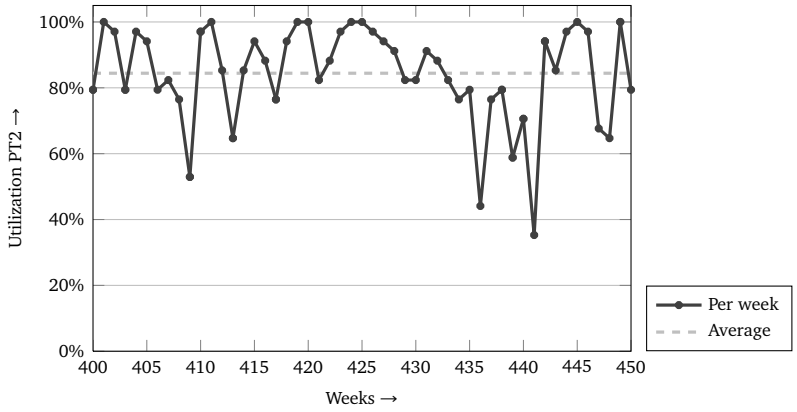


Figure 5.5: Utilization of Physiotherapist 2 during 50 weeks of a simulation run with a total length of 12 years (Scenario 3).

would presumably be present on most weekdays, there would be more possibilities for combination appointments. In addition, a clinic with a larger number of both therapists and patients would be less sensitive to demand fluctuations. Hence, we believe that, due to economies of scale, the potential of our approach for larger clinics is even greater than demonstrated in this chapter.

Given the results of the AMC case, we are convinced that this methodology can be valuable to many rehabilitation outpatient clinics on the operational, tactical, and strategic planning levels. On the operational level, the ILP can be used for scheduling appointments. This process would require customization of the methodology to match the specific restrictions and preferences of each particular clinic. This customization is certainly possible as the ILP approach is suitable for changing or adding constraints and modifying the objective function. On the tactical level, by simulating the application of the methodology, therapist agendas can be aligned. The ILP method can also be beneficial on a strategic planning level, to rationalize the planning strategy and to expose the influence of increasing the relative importance of a particular performance indicator on overall performance. Moreover, the effects of changes in the case mix can be investigated, and insight can be acquired in rationally determining the relative capacities per discipline.

In follow-up research, we focus on three directions. First, as mentioned in Section 5.4.3, reserving capacity for both future patients and patients already under treatment might be a possibility to keep achieving excellent scores for all performance indicators under a high therapist load. Second, in our experiments we have observed substantial variability in therapist utilization from week to week (see Figure 5.5). Balancing out of the utilization per therapist may be favorable. This balancing may possibly be achieved by taking the current utilization of therapists into account when assigning new patients to therapists. Third, as pointed out before, balancing the capacities of the various disciplines is of utmost importance. It may improve the performance of the system as a whole because it may positively affect all performance indicators. As aligning these capacities is not trivial due to the interactions between the disciplines, this area is an interesting direction for future research. This will be the focus of the next chapter.

To conclude, this study demonstrated that the worldwide organizational challenges recently established by the WHO can be well addressed by exploiting operations research techniques. Bringing together healthcare professionals and operations researchers can result in considerable improvements in both service quality and patient-centeredness for the rehabilitation sector.

5.6 Appendix

This appendix contains the mathematical formulation of the ILP. Tables 5.6, 5.7, and 5.8 provide a summary of the notation used. The presented formulation of the ILP is not entirely linear, but linearization is straightforward and is performed automatically by ILOG OPL, in which the ILP was implemented.

Decision variables

We use index a for appointments, h for therapists, and t for time slots (see also Table 5.6). Each day is divided into D time slots. Time slots are numbered consecutively, so $t = 1$ is the first time slot on day one, $t = D + 1$ is the first time slot on day two, and so on. We use the notation $\mathcal{T}_d$ for the set of time slots on day d and $\mathcal{T}_w$ for the set of time slots in week w.

For each appointment within a series, we must select the therapist to whom the patient is assigned and the starting time slot. Hence, the decision variables are as follows:

$$x_{aht} = \begin{cases} 1 & \text{, if appointment } a \text{ is assigned to therapist } h \\ & \text{ and starts in time slot } t, \\ 0 & \text{, otherwise.} \end{cases}$$

To limit computation time, we do not construct decision variables x_{aht} that are not allowed. That is, x_{aht} is not constructed in the following cases:

- the disciplines of appointment a and therapist h do not match
- therapist h is not available in time slot t
- the patient is not available in time slot t
- time slot t is too near to the end of a day, such that appointment a could not be finished before the end of the day if it were started in time slot t
- the patient is not treated by therapist h (only applicable to patients who have already had treatments)

Constraints

In this section, we present the constraints of the model. Several types of constraints are considered. In addition to basic planning constraints, we distinguish constraints with respect to unscheduled appointments, therapist assignment, number of appointments per period, start and continuity of the rehabilitation process, patient preferences, recurring day and time, and the efficient filling of therapist schedules.

Basic planning constraints. Let M_a be the duration of appointment a. Any two appointments of the patient may not overlap. Starting with appointment a, other appointments $\hat{a}$ may not start at time slots in which appointment a is taking place:

$$\sum_{\hat{h},\hat{a}\neq a} x_{\hat{a}\hat{h}\hat{t}} + x_{aht} \leq 1 \quad \text{, for all } a,h,t,\hat{t} | t \leq \hat{t} \leq t + M_a - 1. \tag{5.1}$$

An appointment may only be scheduled if both the patient and the therapist are available. Let G_{ht} be 1 if therapist h is available in time slot t, and let H_t be 1 if the patient is available in time slot t. Thus, we have to require the following:

$$x_{aht} \leq G_{h\hat{t}} \cdot H_{\hat{t}} \quad \text{, for all } a,h,t,\hat{t} | t \leq \hat{t} \leq t + M_a - 1. \tag{5.2}$$

Table 5.6: Indices and sets ILP.

Index	*Description*	*Set*	*Description*
$t, \hat{t}$	time slots	$\mathcal{T}_d$	time slots on day d
d	days	$\mathcal{T}_w$	time slots in week w
w	weeks	$\mathcal{D}_{Y_a}$	days in the week before week Y_a
$h, \hat{h}$	therapists	$\mathcal{D}_{Z_a}$	days in the week after week Z_a
c	disciplines		
$a, \hat{a}$	appointments		

The treatment plan may contain precedence relations between certain appointments. Let parameter $B_{a\hat{a}}$ be 1 if appointment a should take place before $\hat{a}$ and 0 otherwise. To satisfy the precedence relations, we have to require the following:

$$\sum_{\hat{t} \leq t} B_{a\hat{a}} \cdot x_{\hat{a}\hat{h}\hat{t}} \leq 1 - x_{aht} \quad \text{, for all } a, \hat{a}, h, \hat{h}, t. \tag{5.3}$$

Unscheduled appointments. As pointed out in Section 5.3.2, we allow a limited number of unscheduled appointments. The variable n_a is 1 if appointment a is not scheduled and 0 otherwise:

$$\sum_{h,t} x_{aht} = 1 - n_a \quad \text{, for all } a. \tag{5.4}$$

As it is undesirable to omit appointments, the number of unscheduled appointments is penalized in the objective function. For each discipline c, the number of unscheduled appointments is limited to a maximum of 1 in every R appointments that are prescribed in the treatment plan. Recall that when scheduling a series of appointments for a patient, previous series of appointments may already have been scheduled for this patient in the past. Let P_c be the number of appointments prescribed for discipline c in previous series, Q_c the number of those appointments that have not been scheduled, and O_c the number of appointments prescribed in the current series. Furthermore, I_{ac} is 1 if appointment a belongs to discipline c and 0 otherwise. Thus, for the limitation on the number of unscheduled appointments per discipline, we have the following:

$$Q_c + \sum_a I_{ac} \cdot n_a \leq \frac{1}{R}(P_c + O_c) \quad \text{, for all } c. \tag{5.5}$$

Therapist assignment. For each discipline, all appointments have to be assigned to the same therapist. This so-called longitudinal continuity of care is a means of improving patient satisfaction and outcomes of care [34]. We introduce the auxiliary variables y_h that equal 1 if the patient is assigned to therapist h and 0 otherwise:

$$x_{aht} \leq y_h \quad \text{, for all } a, h, t. \tag{5.6}$$

Table 5.7: Parameters ILP.

Parameters	*Description*
Binary parameters	
G_{ht}	1 if therapist h is available in time slot t
H_t	1 if the patient is available in time slot t
$B_{a\hat{a}}$	1 if appointment a should take place before $\hat{a}$
I_{ac}	1 if appointment a belongs to discipline c
J_{hc}	1 if therapist h belongs to discipline c
A_{ht}	1 if an appointment of the previously scheduled series of the patient is assigned to therapist h and starts in time slot t
F_{ac}	1 if appointment a is the first appointment for discipline c according to the treatment plan
N	1 if the patient is a new patient
$E_{t\hat{t}}$	1 if time slot t and $\hat{t}$ are on the same day
General integer parameters	
D	number of time slots per day
M_a	duration of appointment a
R	number of appointments per discipline, of which at most one may be unscheduled
P_c	number of appointments prescribed for discipline c in previous series is exceeded by more than two weeks
Q_c	number of appointments prescribed but not scheduled for discipline c in previous series
O_c	number of appointments prescribed for discipline c in the current series
L	maximum allowed number of appointments with one therapist in a week
K	maximum allowed number of appointments on a single day
S	preferred maximal access time (# weeks)
W	number of time slots per week
C	factor by which the exceeding of the access time is limited
V	number of days within which all first appointments preferably take place (simultaneous start)
Y_a	number of the first week in which appointment a may be scheduled
Z_a	number of the final week in which appointment a may be scheduled
Φ	number of days that have passed since the start of the treatment
T	number of time slots in planning horizon
Θ	number of weeks delay in treatment process
Ψ	prescribed duration of series of appointments
Ω	minimal number of appointment days required
U	maximum allowed waiting time for the patient between two consecutive appointments

Let parameter J_{hc} be 1 if therapist h belongs to discipline c. We enforce longitudinal continuity of care by the following equation:

$$\sum_{h} J_{hc} \cdot y_h \leq 1 \quad \text{, for all } c. \tag{5.7}$$

By not constructing decision variables $x_{a\hat{h}t}$ for therapists $\hat{h}$ who do not treat the patient, we will require that $y_h = 1$ if the patient has had treatments from therapist h in previous series.

Number of appointments per period. Multiple appointments with the same therapist may not be scheduled on the same day d. Let A_{ht} be 1 if an appointment of the previously scheduled series of the patient is assigned to therapist h and starts in

Table 5.8: Variables ILP.

Variables	*Description*
Binary variables	
x_{aht}	1 if appointment a is assigned to therapist h and starts in time slot t
n_a	1 if appointment a is not scheduled
y_h	1 if the patient is assigned to therapist h
s_a	1 if appointment a takes place one day after a previous appointment with the same therapist
e_d	1 if appointments for the patient are scheduled on day d
m	1 if the patient has no simultaneous start with the various disciplines
q_a	1 if appointment a may not be scheduled a week earlier than prescribed in the treatment plan
r_a	1 if appointment a may not be scheduled a week later than prescribed in the treatment plan
z_1	1 if prescribed duration of the series of appointments is exceeded by two weeks or less
z_2	1 if exceeding of prescribed duration of series of appointments is between one and two weeks
z_3	1 if prescribed duration of series of appointments is exceeded by more than two weeks
τ_t	1 if t is a non-recurring starting time slot
i_a	1 if appointment a causes idle time in the schedule of the therapist beforehand
j_a	1 if appointment a causes idle time in the schedule of the therapist afterwards
g_a	1 if appointment a causes idle time in the schedule of the therapist both beforehand and afterwards
General integer variables	
f	number of the starting time slot of the first appointment
k	number of the day on which the first appointment is scheduled
b	number of time slots by which the preferred access time is exceeded
u_a	number of time slots that appointment a is scheduled before week Y_a
v_a	number of time slots that appointment a is scheduled after week Z_a
p	difference between the number of appointment days realized and Ω
μ	excess number of non-recurring starting time slots

time slot t. Recall that $\mathcal{T}_d$ denotes the set of time slots on day d. Then, we require the following:

$$\sum_{t \in \mathcal{T}_d} \left(A_{ht} + \sum_{a} x_{aht} \right) \leq 1 \quad \text{, for all } h, d. \tag{5.8}$$

Preferably, multiple appointments with one therapist are evenly spread over a week. Hence, we will penalize situations in which appointments with one therapist are scheduled on consecutive days. Let s_a be 1 if appointment a is scheduled such that it takes place one day after a previous appointment with the same therapist. We penalize s_a in the objective function. Let d_1 denote the day after the day of time slot t. Therefore, the constraint is as follows:

$$\sum_{\hat{t} \in \mathcal{T}_{d_1}} \left(A_{h\hat{t}} + \sum_{\bar{a}} x_{\bar{a}h\hat{t}} \right) + x_{aht} \leq 1 + s_a \quad \text{, for all } a, h, t. \tag{5.9}$$

To also enhance the spreading out of the treatments per discipline over weeks, the number of appointments with one therapist in a week is limited to L. Remember that $\mathcal{T}_w$ denotes the set of time slots in week w. Hence, the constraint is as follows:

$$\sum_{t \in \mathcal{T}_w} \left(A_{ht} + \sum_{a} x_{aht} \right) \leq L \quad \text{, for all } h, w. \tag{5.10}$$

As treatments may be strenuous for the patient, the number of appointments that may be scheduled on a single day is limited to K. We introduce auxiliary variables e_d which are 1 if one or more appointmentsare scheduled on day d and 0 otherwise:

$$\sum_{t \in \mathcal{T}_d} \sum_{h} \left(A_{ht} + \sum_{a} x_{aht} \right) \leq K \cdot e_d \quad \text{, for all } d. \tag{5.11}$$

Start of the rehabilitation process. As we want to control the access time, we have to identify the number f of the starting time slot of the very first appointment. Let parameter F_{ac} be 1 if appointment a is the first appointment for discipline c according to the treatment plan and 0 otherwise. Then, we obtain the following:

$$f = \min_{c} \left\{ \sum_{a,h,t} (F_{ac} \cdot t \cdot x_{aht}) \right\}. \tag{5.12}$$

Based on f, the number k of the day on which the very first appointment takes place is as follows:

$$k = \left\lceil \frac{1}{D} \cdot f \right\rceil. \tag{5.13}$$

The access time of the patient should preferably be within S weeks. Let W be the number of time slots in a week and N be 1 if the patient is a new patient and 0 otherwise. We introduce the variable b, which is the number of time slots by which the access time exceeds the preferred access time ($b \geq 0$):

$$N \cdot (f - b) \leq S \cdot W. \tag{5.14}$$

We limit exceeding of the access time by requiring that b may be no larger than C times the preferred access time:

$$b \leq C \cdot S \cdot W. \tag{5.15}$$

Patients who cannot be seen within the preferred access time plus the maximum allowed extension, are instead referred to another rehabilitation clinic, as clinicians indicate that quality of care cannot be guaranteed when the access time exceeds this threshold.

For the rehabilitation process it is preferable that the patient starts treatment with all of the various relevant disciplines simultaneously. Therefore, we would like the first appointment with each discipline to take place within V days of the very first appointment. We introduce the variable m, which is 1 if this preference is not satisfied, and penalize m in the objective function:

$$N \cdot \sum_{a,h,t} (F_{ac} \cdot t \cdot x_{aht}) \leq D \cdot (k + V - 1) + W \cdot m \quad \text{, for all } c. \tag{5.16}$$

Continuity of the rehabilitation process. For each appointment a, the treatment plan prescribes the range of weeks within which it should be scheduled (counting

from the week in which the rehabilitation process started). Let Y_a be the number of the first week in which a may scheduled and Z_a be the number of the final week. Now we would like to schedule a in one of the weeks $Y_a, \ldots, Z_a$. As a deviation from these preferred weeks is better than not scheduling a at all, we allow for some (penalized) scheduling flexibility: a may be scheduled a week earlier than week Y_a or a week later than week Z_a if the patient does not already have an appointment with that same discipline during these other weeks. Hence, we first determine whether or not this situation applies. We introduce variables q_a (r_a), which are 1 if appointment a may not be scheduled a week earlier (later) and 0 otherwise. Let $\mathcal{D}_{Y_a}$ denote the set of days in the week before week Y_a. Thus, we require the following:

$$\sum_{h,c,t \in \mathcal{T}_d} \sum_{d \in \mathcal{D}_{Y_a}} \left(A_{ht} + \sum_{\hat{a}} x_{\hat{a}ht} \right) \cdot I_{ac} \cdot J_{hc} \leq L \cdot q_a \quad \text{, for all } a. \tag{5.17}$$

Similarly, if $\mathcal{D}_{Z_a}$ denotes the set of days in the week after week Z_a, we need the following:

$$\sum_{h,c,t \in \mathcal{T}_d} \sum_{d \in \mathcal{D}_{Z_a}} \left(A_{ht} + \sum_{\hat{a}} x_{\hat{a}ht} \right) \cdot I_{ac} \cdot J_{hc} \leq L \cdot r_a \quad \text{, for all } a. \tag{5.18}$$

In case appointment a has to be scheduled before week Y_a, the variable u_a counts the number of time slots between the start of a and the start of week Y_a. Now, u_a may be at most a week, unless a may not be scheduled earlier:

$$u_a \leq W - W \cdot q_a \quad \text{, for all } a. \tag{5.19}$$

In case appointment a has to be scheduled after week Z_a, the variable v_a counts the number of time slots between the end of week Z_a and the start of a, and we require the following:

$$v_a \leq W - W \cdot r_a \quad \text{, for all } a. \tag{5.20}$$

Now, we would like to schedule each appointment a in the week or range of weeks prescribed in the treatment plan or set u_a (v_a) to the right value if a is scheduled earlier (later) than prescribed. In the latter case, we penalize for this in the objective function. If a can neither be scheduled in the prescribed weeks nor earlier or later, a is not scheduled at all, and n_a is set to 1. Let T be the total number of time slots in the planning horizon, Φ the number of days that have passed since the start of the rehabilitation process, and Θ the number of week-long delays since the start of the rehabilitation process. To not schedule a too early, we require the following:

$$1 + N \cdot D \cdot (k-1) + W \cdot (\Theta + Y_a - 1) - u_a \leq D \cdot \Phi + \sum_{h,t} t \cdot x_{aht} + T \cdot n_a \quad \text{, for all } a. \tag{5.21}$$

Similarly, to not schedule a too late, we require the following:

$$D \cdot \Phi + \sum_{h,t} t \cdot x_{aht} \leq N \cdot D \cdot (k-1) + W \cdot (\Theta + Z_a) + v_a \quad \text{, for all } a. \tag{5.22}$$

The lead time of the rehabilitation process, from the first until the last appointment, should preferably be as prescribed in the treatment plan. It is undesirable to lengthen the lead time for scheduling reasons. Let Ψ be the prescribed duration in weeks of a series of appointments. We introduce the variables z_1, z_2, and z_3. If the prescribed duration is exceeded by one week or less, z_1 is 1. Otherwise, if the actual duration exceeds the prescribed duration by between one and two weeks, both z_1 and z_2 are 1. If the duration exceeds the prescribed length by more than two weeks, see the following z_3 is 1:

$$\max_{a,h,t}\{t \cdot x_{aht}\} - N \cdot f \leq W \cdot \big(\Psi + z_1 + z_2 + T \cdot z_3\big). \tag{5.23}$$

When the prescribed duration is exceeded, this is penalized with the weights γ_1, γ_2, and γ_3 (for z_1, z_2, and z_3, respectively), where $\gamma_1 < \gamma_2$ and $\gamma_1 + \gamma_2 < \gamma_3$.

Patient preferences. Combination appointments are high on the list of outpatient preferences [603]. Therefore, we strive to schedule the appointments on as few days as possible. We introduce a parameter Ω representing the minimal number of 'appointment days' required given the constraints of no more than K appointments on a single day (5.11) and that multiple appointments with the same therapist may not be scheduled for the same day (5.8). The variable p that is penalized in the objective function represents the difference between the true number of 'appointment days' and Ω ($p \geq 0$):

$$\sum_{d} e_d \leq \Omega + p. \tag{5.24}$$

To limit patients' waiting time between appointments on the same day, these time intervals between two consecutive appointments in one day should not exceed U time slots. Let $E_{t\hat{t}}$ be 1 if time slots t and $\hat{t}$ fall on the same day. Thus, we have to require the following, for all a, h, t:

$$\sum_{\hat{t}=t+M_a+U+1}^{t+D} E_{t\hat{t}} \cdot \sum_{\hat{h}} \Big(A_{\hat{h}\hat{t}} + \sum_{\hat{a}} x_{\hat{a}\hat{h}\hat{t}}\Big) \leq$$
$$K \cdot \sum_{\hat{t}=t+M_a}^{t+M_a+U} E_{t\hat{t}} \cdot \sum_{\hat{h}} \Big(A_{\hat{h}\hat{t}} + \sum_{\hat{a}} x_{\hat{a}\hat{h}\hat{t}}\Big) + K \cdot (1 - x_{aht}) \tag{5.25}$$

Recurring day and time. It is preferred that the appointments take place on the same day and time each week, such that the patient has to remember only a short list of days and times. Hence, for an appointment that starts in time slot t, we first verify whether or not another appointment has been scheduled for the same time

slot in one of the previous weeks (i.e., a multiple of W time slots before t). If not, the binary variable τ_t is set to 1, indicating that t is a non-recurring appointment time slot:

$$x_{aht} - \sum_{w|w\cdot W<t} \sum_{\hat{h}} \left(A_{\hat{h}(t-w\cdot W)} + \sum_{\hat{a}} x_{\hat{a}\hat{h}(t-w\cdot W)} \right) \leq \tau_t \quad \text{, for all } a,h,t. \tag{5.26}$$

Clearly, the number of non-recurring appointment time slots is at least equal to the maximum number of appointments that take place within one week. We let the variable μ count and penalize the excess non-recurring appointment time slots by adding μ to the objective function. The constraint is as follows:

$$\sum_t \tau_t - \mu \leq \max_w \Big\{ \sum_{t\in\mathcal{T}_w} \sum_h \Big(A_{ht} + \sum_a x_{aht} \Big) \Big\}. \tag{5.27}$$

Efficient filling of therapist schedules. For the convenience of the therapists and to achieve a high utilization rate, it is preferable to avoid idle time in the schedules of therapists between two consecutive appointments in a day. As it might be impossible to later fit another appointment into this idle time, the prevention of idle time minimizes the number of referred patients and unscheduled appointments. Hence, we aim to schedule appointments right at the start or at the end of a session of the therapist, or right before or after an already scheduled appointment. We introduce the variable i_a (j_a), which is 1 if appointment a is scheduled in such a way that idle time is caused in the schedule of the therapist before (after) a. Then, we have to require the following:

$$i_a \geq G_{h(t-1)} \cdot x_{aht} \quad \text{, for all } a,h,t, \tag{5.28}$$

$$j_a \geq G_{h(t+M_a)} \cdot x_{aht} \quad \text{, for all } a,h,t. \tag{5.29}$$

If an appointment a is scheduled in such a way that it causes idle time in the schedule of the therapist both beforehand and afterwards, we say that a causes a break in the schedule of the therapist. This break is penalized in the objective function by the variable g_a, which is 1 in this case (and 0 otherwise):

$$i_a + j_a \leq 1 + g_a \quad \text{, for all } a. \tag{5.30}$$

Objective function

The objective function of the model, of which the components were presented in detail in Section 5.3.2, is as follows:

$$\min \Bigg\{ \alpha \cdot \left\lceil \frac{b}{D} \right\rceil + \beta \cdot m + \sum_{i=1}^{3} \gamma_i \cdot z_i + \delta \cdot p + \zeta \cdot \sum_a g_a + \eta \cdot \sum_a n_a + \theta \cdot \sum_a \left\lceil \frac{u_a + v_a}{D} \right\rceil + \kappa \cdot \sum_a s_a + \chi \cdot \mu \Bigg\}.$$

CHAPTER 6

Balancing Discipline Capacities

6.1 Introduction

In this chapter, we perform a patient flow analysis of the Dutch rehabilitation center 'Het Roessingh' and address the related resource capacity planning and control issues. In particular, it connects with the challenge that was stated at the end of the previous chapter by focusing on the capacity dimensioning of the various involved disciplines in rehabilitation care. Rehabilitation care is the process in which a patient is assisted in improving or recovering lost functions after an event, illness or injury that causes functional limitations. The patient is treated by a multidisciplinary team of a rehabilitation physician and therapists for a period of time. Each team member treats the patient in different segments of the rehabilitation process.

Het Roessingh provides a striking example of the large potential for organizing rehabilitation care more efficiently and effectively that is observed by both the WHO [640] and the Dutch rehabilitation sector [522]. While the treatments at its Pain Rehabilitation department are effective (the treatment programs are accredited by the Commission on Accreditation of Rehabilitation Facilities [131]), its management observes considerable organizational challenges: waiting lists are long (on average patients wait more than 80 days for their first consultation), practitioners experience high working pressure, and the insight into the demand and supply of pain rehabilitation care is insufficient.

To tackle these challenges, inspired by the concept of clinical pathways [148], Het Roessingh is introducing the concept of 'treatment plans'. Treatment plans specify the required treatment for specific groups of patients with the same diagnosis during a period of several weeks or months. They intend to ensure that the right treatments are provided at the right time, so that the best quality of care is realized while making efficient use of available resources [144].

Treatment plans resemble the well-known clinical or critical pathways but are in fact not the same. Although a unique definition of clinical pathways is lacking [597], the definition given by Medical Subject Headings (MeSH) of Pubmed [574] is: "schedules of medical and nursing procedures, including diagnostic tests, medications, and consultations designed to effect an efficient, coordinated program of treatment." A treatment plan at Het Roessingh differs from a clinical pathway as is it used as a blueprint for a patient's treatment for some weeks or even months (like

in Chapter 5), where a clinical pathway typically prescribes in detail all required activities and their timing during a time horizon of a number of days. A thorough description of clinical pathways, their creation and their usage can be found in [95]. A literature review on the usage of clinical pathways in clarifying patient flows is given in [591].

The introduction of treatment plans has both a medical and a logistical motivation, as it creates clarity for both patients and practitioners. It realizes uniformity in the care process, so that the risk of both undertreatment and overtreatment are minimized [522]. Also, it prevents discontinuity in the care process and it stimulates the coordination among the different practitioners in question. From a planning perspective, the treatment plan concept offers insight in required capacity for the patient mix the facility serves. In addition, instead of scheduling the treatment of a patient one week at a time, a treatment plan offers the opportunity to schedule the total treatment of a patient at once.

The complexity of a rehabilitation care chain is induced by its multidisciplinarity and length of the treatments, which results in logistical interactions between the care providers. The methods and analyzes presented in this chapter support Het Roessingh to make their organizational process ready for complete implementation of treatment plans. They support Het Roessingh to gain insight into the behavior of their care chain, and to obtain insight in demand for care and the capacity required to meet a certain quality of service. The integral patient flow is addressed, from the intake consultation until the end of treatment. The analysis enables us to advise Het Roessingh on the optimal system configuration of the treatment plan based care chain and to derive rules of thumb that can be applied in its design and control.

In Chapter 5, we presented an algorithm to schedule an entire multidisciplinary rehabilitation treatment at the moment a patient sends in an application. The discipline capacities were taken as a given. Determining the staff capacity dimensioning, which includes the size of the workforce and its skill mix, is an important aspect in healthcare as can be seen by the many literature reviews on this topic [90, 142]. For relevant OR/MS references on staff capacity dimensioning, we refer the reader to Chapter 2. To best of our knowledge, no references are available on supporting decision making with respect to staff capacity dimensioning for multidisciplinary treatment plans.

The chapter is organized as follows. Section 6.2 introduces the case study of Het Roessingh, discusses their main organization challenges, and provides a detailed specification of our contributions. Section 6.3 describes the developed OR/MS methods, which are based on queueing theory, Markov models, and discrete-event simulations. Section 6.4 presents the numerical results and our main insights. The chapter is closed with a discussion of our findings in Section 6.5.

6.2 Background: case study

Het Roessingh employs over 600 persons (400 FTE) performing rehabilitation treatment and medical research. Each year, about 3800 patients are treated in this center

in three different departments: Adult Rehabilitation, Children's Rehabilitation and Pain Rehabilitation.

In this study we focus on the Pain Rehabilitation department. This department treats over 900 patients per year (roughly 53.000 consultations) suffering from chronic pain or chronic fatigue syndrome for which no medical treatment is known. Patients are taught to cope with their pain, thereby trying to enable them to fully participate in society. Clinicians from six medical disciplines are involved in the treatment at Pain Rehabilitation: Rehabilitation Physicians (RP), Physiotherapists (PT), Occupational Therapists (OT), Psychologists (PS), Social Workers (SW), and Kinesiologists (KI).

The current patient flow, before the introduction of treatment plans, is depicted in Figure 6.1. Patients who are referred to the pain rehabilitation department (17.7 per week on average) by a physician, will generally first receive an intake consultation (93.9%). The intake consists of interviews by multiple therapists of different disciplines and a rehabilitation physician. Advised by the multidisciplinary team of therapists, the physician decides whether the patient is eligible for treatment at Het Roessingh (81.6% of the intake patients) and if so which treatment is most suitable. If no proper diagnosis could be made during the intake, the first period of the treatment consists of an 'observation period' (14.6% of the intake patients), during which the appropriate follow-up treatment is determined. Two types of patients do not require an intake before their treatment, because their medical condition is already known in detail: patients who receive rehabilitation treatment after a cancer treatment (5%) and patients that require only a specific physiotherapy treatment (1.1%).

Patients are treated as (semi-)inpatient (62%) or outpatient (38%). The (semi-) inpatients require a bed during (part of) the weeks they are in treatment. The majority of the treatments are organized in groups of 6 to 10 patients. Periodically, the rehabilitation physician and the therapists discuss the patient's condition during an multidisciplinary team meeting (MTM), to decide whether the treatment should be continued, adjusted or finished.

Improvement potential can be identified in three main drivers behind the rehabilitation care organization of Het Roessingh: coordination, clinician load and capacity balancing.

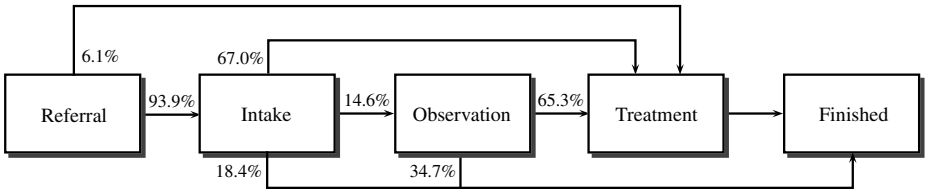


Figure 6.1: Patient flow diagram in the Pain Rehabilitation department of Het Roessingh.

Coordination. In current practice, treatment planning is decentralized, which implies that all disciplines or even therapists manage their own agendas. This hampers coordination within both the care process and the logistical process. As a consequence, in many cases, a short access time and a simultaneously started treatment cannot be realized. Also, timely planning of follow-up appointments can be problematic, causing discontinuity of the rehabilitation process. Consequently, certain prescribed treatments may never be realized, since they cannot be scheduled. Besides, without centralized planning, the risk of both undertreatment and overtreatment is observed and management experiences difficulties in effectively controlling the organization.

Clinician load. Several disciplines experience demand overload. As a result, waiting lists are long (months) and growing, work pressure is high, and the total duration of treatments is unpredictable due to the occurring discontinuity. The management indicates that therapists do not have sufficient time for administrative duties related to a patient's treatment, so-called indirect patient care. This has a negative effect on the reporting quality, which results in the cancelation of MTMs where a patient can be discharged, since these can only effectively be performed if the required documentation is available. If an MTM is canceled, in general, the treatment is extended at least until the next MTM. This effect fortifies itself, since unnecessarily extending treatments claims capacity which could otherwise be used to admit new patients from the waiting list.

Capacity planning. Measurements reveal that there exists an imbalance between the capacities that the different disciplines have available. Consequently, while some disciplines experience demand overload, others experience underutilization. This results in a situation where waiting lists are growing, while at the same time available capacity remains unused.

The intervention that Het Roessingh intends to make to address the coordination issue is to introduce treatment plans. Each treatment plan takes a fixed number of weeks. The entire treatment is scheduled in advance, including the MTMs, in which the progress of the patient is discussed. Since a simultaneous start at the different disciplines is essential for a coordinated care process, a patient's treatment is then only allowed to start if each of the required disciplines has capacity available for the complete period of the treatment.

Here, we study the effect of the introduction of treatment plans on clinician load and capacity balancing issues. In more detail, our contributions are as follows: (i) mapping of the pain rehabilitation process and treatment plans (Section 6.3.1), (ii) determining the capacity required for intake consultations (Sections 6.3.2 and 6.4.1), (iii) illustrating the impact of canceling MTMs (Sections 6.3.3 and 6.4.2), (iv) computing the mean and standard deviation of the demand per discipline (Sections 6.3.4 and 6.4.3), (v) defining rules of thumb for balancing discipline capacities (Sections 6.3.5 and 6.4.4), and (vi) evaluating the complete system redesign in terms of quality of service and logistical efficiency (Sections 6.3.6 and 6.4.4).

6.3 Methods

There are in total 18 treatment plans which cover 6 disciplines and may last up to 39 weeks, which makes the system under study complex. For data collection and process mapping we have used the data systems in Het Roessingh and interviews with the staff including doctors, managers and IT specialists. An important attribute of our approach is *decomposition* of the process into simpler processes where possible. In particular, we develop three mathematical models that address two specific aspects of the patient flows that can be studied independently. (1) Intakes happen at the beginning of the treatment plan and do not depend on the future treatment. Therefore, we analyze the required number of intakes per week with a separate queueing model. (2) Cancelation of MTMs, where patients are discussed and discharged, result in sometimes unnecessarily prolonged treatment plans. We investigate the effect of such prolongation for each treatment plan separately. (3) We develop a Markov model that describes the progress of patients through treatment plans. The proposed analytic models are then combined in a simulation model. In this section, we describe the methods in more detail.

6.3.1 Process mapping and patient flow model

First, we estimate patients' pathways through (usually multiple) treatment plans. The treatment plans created by Het Roessingh are blueprints. Some treatments in treatment plans are given only to a fraction of patients, and the duration of the appointments can vary from one patient to another.

We have used historic data and expert opinion to obtain the required information. This resulted in 18 treatment plan blueprints (examples are oncology, whiplash, and back pain). Each treatment plan blueprint prescribes the disciplines a patient of a particular type should be treated by, the required number of treatments per discipline, the duration of each treatment (in minutes) and the week number in which it should take place. In addition, the fractions are included according to which particular appointments series are required for a specific patient, so that individualized treatment plans can be realized based on the treatment plan blueprints. The results for one of the two observation treatment plans are presented in Table 6.1. This treatment plan of five weeks involves six disciplines.

Next, we analyzed *sequential treatment plans and transition probabilities.* Usually a treatment plan is only a part of a complete care process. For example, 'Intake' and 'Observation' are treatment plans, which are usually followed by other treatment plans. Thus, we need to determine transition probabilities between treatment plans. For instance, if the transition probability from treatment plan 1 to treatment plan 2 is 0.15, then 15% of all patients who completed treatment plan 1 will continue with treatment plan 2. Since treatment plans are not yet applied in practice, the transition probabilities have been estimated through historic data on the current process. The transition probabilities were recorded in matrix Q, where entry (j,k), $Q_{(j,k)}$, is the probability that treatment plan j is followed by treatment plan k.

Table 6.1: Example of a treatment plan blueprint (with the treatment durations in minutes).

Discipline	*Appointment*	*Group size*	*Treatment duration in week 1*	*2*	*3*	*4*	*5*	*Required for x%*
Rehabilitation physicians (RP)	Therapy 1	1	20	20	20	-	-	100
	Therapy 2	1	-	-	-	-	30	100
	MTM	1	15	15	15	15	-	100
Physiotherapy (PT)	Therapy 1	1	60	60	60	-	-	100
	MTM	1	15	15	15	15	-	100
Occupational therapy (OT)	Therapy 1	1	60	60	60	-	-	75
	MTM	1	15	15	15	15	-	100
Psychology (PS)	Therapy 1	1	60	60	60	-	-	100
	Therapy 2	1	60	60	-	-	-	80
	Therapy 3	1	-	-	-	-	30	100
	MTM	1	15	15	15	15	-	100
Social workers (SW)	Therapy 1	1	60	60	60	-	-	80
	Therapy 2	1	60	-	-	-	-	100
	MTM	1	15	15	15	15	-	100
Kinesiology (KI)	Therapy 1	3	30	30	30	-	-	100
	Therapy 2	3	30	30	30	-	-	100
	Therapy 3	1	30	30	30	-	-	75
	MTM	1	15	15	15	15	-	100

Combining the obtained information on treatment plans and their sequence, the following Markov model describes the possible treatment scenarios of a patient. Let l_j be the duration in weeks of treatment plan $j = 1, 2, \ldots, 18$. For each week number $w = 1, 2, \ldots, l_j$ in a treatment plan j, and for each discipline $d = 1, 2, \ldots, 10$, we are given a complete description of the number of appointments of type a, denoted by $n_{w,j}^{d,a}$, and the probability that an appointment is required, denoted by $p_{w,j}^{d,a}$. After completion of a treatment plan j, we assume that transitions between treatment plans happen according to a Markov process, that is, a patient proceeds to treatment plan k with probability $Q_{(j,k)}$, independent of the past.

Finally, we establish *priority rules*. By arrival, after intake, and between treatment plans patients can be put on a waiting list if capacity is insufficient to treat them immediately. Priority rules determine which patients will be treated first if capacity is insufficient to treat all patients on the waiting list. Het Roessingh uses the following order of priority: (1) intake; (2) observation; (3) group treatments; (4) individual treatments.

6.3.2 Number of intakes

Intakes take place for multiple patients simultaneously because intakes usually require similar sets of specialists and procedures. That is, each week K intake sessions are held, each consisting of N patients. Intake sessions take place only when they are completely filled; if the number of patients waiting for intake is less than N

the intake session is canceled. Denote the average number of new patients registered in one week by λ. Then the questions that are of interest for Het Roessingh can be formulated as follows. Given λ, K and N, what is the average access time for an intake and what is the probability that the access time exceeds a specified norm? We answer these questions using a queueing model, which is presented in Appendix 6.6.1.

6.3.3 Multidisciplinary team meetings

At fixed points in time, at the MTMs for individual outpatients, specialists discuss the progress of a patient and the continuation of the treatment. The first meeting takes place after three weeks, and subsequently every sixth week. When such a meeting is canceled then usually the treatment period of a patient is extended. Practice shows that the last meeting in the treatment plan is crucial, and whenever this meeting is canceled the treatment will definitely be extended. For any other meeting the choice of extending the treatment depends highly on the patient's condition. When a treatment is extended, it will be extended with one period of six weeks and one MTM.

In practice, there is no formal policy for treatment extensions. We consider two possible situations: (1) each canceled meeting leads to an extension of the treatment; (2) only cancelation of the last meeting leads to an extended treatment. After introduction of the treatment plans, the practical situation will presumably be in between the two options. Applying a simple analytical method (presented in Appendix 6.6.2) reveals the negative effect of canceled meetings on treatment durations. These results are described in Section 6.4.2.

6.3.4 Capacity requirements

We determine the average capacity requirements of the network of patient flows induced by the treatment plans, by (1) calculating the average capacity in hours required per appointment of each type in each week, (2) computing the average capacity required per patient in each week of treatment plan j, and (3) considering all possible sequences of treatment plans $j_1 \to j_2 \to \cdots \to j_k \to j$ that lead to treatment plan j. The network of treatment plans is such that no patient receives a particular treatment plan twice. Combining (1), (2), and (3) with the average number of arrivals per week gives us for each discipline the average capacity required to treat all patients. The mathematical formulation of this method is presented in Appendix 6.6.3.

6.3.5 Capacity dimensioning improvements

In order to balance the load of disciplines and to improve the performance in terms of number of treated patients, we propose to determine the capacity levels such that each discipline has the same (theoretical) load ρ^d, defined as the average capacity requirement for the discipline divided by the available capacity for this discipline.

So, we set the capacities of the disciplines such that the following holds for each discipline d:

$$\rho^d = \mathbb{E}[C^d]/[\text{available capacity of discipline } d] = \rho. \tag{6.1}$$

It is well-known from queueing theory that the load of the system is a defining parameter for the magnitude of queue lengths, access times, and probability of no queue. Therefore, by leveling the load between disciplines, we achieve comparable waiting times for different classes of patients.

Remark 6.1. In the context of call-centers and other applications a so-called square-root staffing rule is often used to define the optimal capacity level S: $S = \lambda + \beta\sigma$, where λ and σ are, respectively, the mean and the standard deviation of the workload offered to the system per time unit. This is the so-called 'square-root staffing rule'. In addition to (6.1), we have also tested such a rule by computing $\sigma[C^d]$ using similar but more involved derivations as in Appendix 6.6.3, and then evaluated capacity levels with different values of β. Since we did not find improvements against the capacity dimensioning rule with equal ρ, we choose to not present the results here. This finding can be explained as follows. The square-root staffing rule is theoretically justified when the amount of work λ grows large, and σ is of the order $\sqrt{\lambda}$, hence, the utilization or load ρ approaches 1 (heavy-traffic regime), see for example [63]. In our situation, however, simulations will show that the utilization of almost 100% cannot be achieved due to the interaction between the disciplines (see Section 6.4.3). Then the square-root staffing rule is not expected outperform the 'equal-ρ rule'.

6.3.6 Simulation model

We investigate the performance under different load and discipline capacity levels using discrete-event simulation of patient flows [383]. The **input** for the simulation model consists of: the treatment plans, the transition matrix Q, the priority rules, the available capacity (either the current capacity or according to our proposed capacity dimensioning rule), and the mean number of weekly referrals. The **process** consists of the discrete-event simulations that mimic, per week, the entire network of patient flows within the Pain Rehabilitation department in Het Roessingh. The number of arrivals in a week is modeled as a Poisson random variable. The simulation model then performs two actions: (1) for each new patient a first treatment plan is determined using the transition matrix Q; (2) each new patient is put on the waiting list corresponding to his/her treatment plan. For each week the simulation model checks whether patients on the waiting lists can start their treatment or not. If there is enough capacity available to treat a patient then the patient starts the treatment, and the available amount of capacity is decreased by the corresponding amount of hours, specified by the treatment plan. Priority rules (see Section 6.3.1) are used to determine the order in which the patients from the waiting list receive their treatment. When a patient has completed his treatment plan then there are two options:

either the patient starts a new treatment plan (which is determined using the transition matrix Q) or the patient leaves the system. The **output** consists of several performance measures: the time a patient must wait before an intake is completed or until the treatment has started, the total duration of a patient's treatment, and the load or utilization of each specific discipline.

6.4 Numerical results

Our models enable Het Roessingh to evaluate various improvements of both quality of care and efficiency, ranging from organizing intake procedures and prioritizing multidisciplinary team meetings, to indicating bottlenecks and determining discipline capacity levels. This section presents the highlights.

6.4.1 Number of intakes

Dutch healthcare organizations have agreed upon an access time norm for rehabilitation care of seeing 80% of the patients within three weeks [568]. In addition, it is a target to see all patients within four weeks. In the current configuration Het Roessingh has $K = 3$ intake sessions of group size $N = 5$, which is insufficient to handle the average number of referrals per week ($\lambda = 17.7$). Our results show that increasing the number of intakes to $K = 3$ and $N = 6$, or $K = 4$ and $N = 5$, will be sufficient to satisfy this norm.

Figure 6.2 gives the average access time for increasing λ, resulting from the queueing model of Section 6.3.2. Observe that two situations lead to extremely long waiting lists and access times. First, when $\lambda \approx K \cdot N$, the load of the system approaches 100%, which results in extremely long access times. Second, when $\lambda \approx 0$ it is difficult to form a group of N patients to plan an intake session because there are almost no arrivals, and thus access times are long. Clearly, the first situation is especially relevant to Het Roessingh. Figure 6.3 displays the access time percentiles for configurations $K = 3$, $N = 6$ and $K = 4$, $N = 5$. The configuration $K = 3$, $N = 6$ is not indicated in Figure 6.3 as it gives 0% seen within four weeks.

For $\lambda = 17.7$, we advise to use $K = 4$ and $N = 5$, and the access time norm is satisfied, all patients are seen within four weeks (>>99.9%), and the utilization of the intake sessions is high: 88.5%.

6.4.2 Multidisciplinary team meetings

Consider a treatment plan of 15 weeks that involves three MTMs. Let p denote the probability of canceled meetings. The average extensions of the treatment duration under the two scenarios that were introduced in Section 6.3.3 are as follows:

- *Scenario 1* (an additional six treatment weeks for every canceled MTM): $\frac{18p}{1-p}$.
- *Scenario 2* (an additional six weeks if the last MTM is canceled): $\frac{6p}{1-p}$.

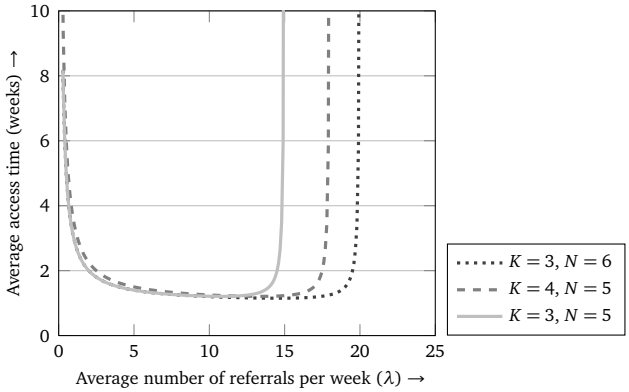


Figure 6.2: The average access time for intakes, depending on λ.

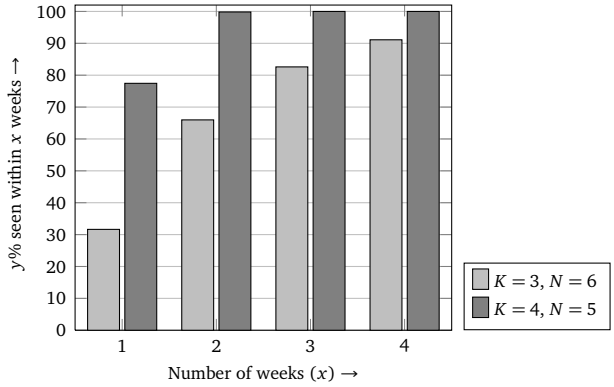


Figure 6.3: Access time percentiles per week for intakes for $\lambda = 17.7$.

The cancelation probability p has a significant effect on the length of the treatment plan, as illustrated in Figure 6.4. For instance, under scenario 2, for $p = 0.1$ the average extension is 0.67 weeks, while for $p = 0.4$ it increases to 4 weeks. Although the clinicians usually give a higher priority to patient treatments than to the MTMs, our results show that the management must facilitate a higher attendance of the meetings, otherwise the scarce capacity will be spend, in large amounts, on patients that could have been discharged, while the patients who need care will be placed on the waiting list.

6.4.3 Capacity requirements and bottlenecks

Evaluating the ability of Het Roessingh to accommodate the implementation of treatment plans with the current discipline capacities, requires the identification of bottlenecks and of the maximum number of patients that can be treated by each discipline.

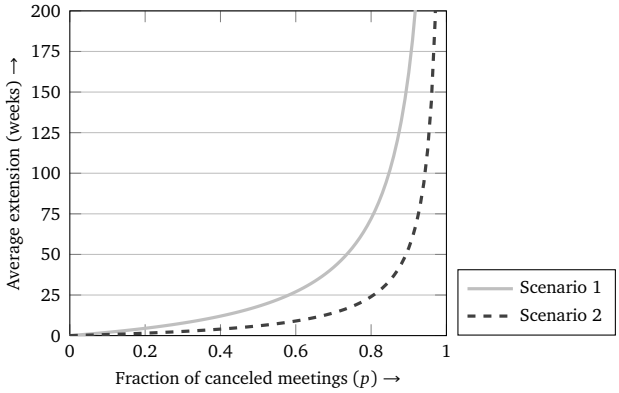


Figure 6.4: An illustration of the effect of canceling MTMs.

The Markov model from Section 6.3.1 in combination with the model from 6.3.4 allows us to identify the bottleneck disciplines, and to determine the maximum number of patient referrals each discipline can handle, by calculating the average capacity requirements per patient referral. The results are presented in Table 6.2. The capacity requirements are measured in Full Time Equivalents (FTEs).

By comparing the capacity requirements to the available capacity for each discipline, we identify the bottlenecks in patient flows. When the average capacity requirement exceeds the available capacity, it is a clear indication that the discipline d constitutes a bottleneck in the system. The fifth column shows that almost all disciplines are overloaded, since the ratio demand over capacity exceeds one. Also, we see that the discipline Occupational Therapists can only handle an average of eight referrals per week. If we compare this with the actual number of referrals per week (which is on average 17.7) it is clear that the discipline OT is the main bottleneck for the network of patient flows induced by the treatment plans. The conclusion is that complete implementation of treatment plans is only possible if a lower number of referrals is allowed or if the capacity levels are adjusted.

6.4.4 Capacity balancing

Het Roessingh strives to uniformize its processes, switch to treatment plans, reduce the work pressure, and avoid ad hoc decisions in the future. Our results help to achieve these goals by having revealed which disciplines in Pain Rehabilitation suffer from the highest overload in the desirable system design, and by proposing capacity dimensioning improvements. To illustrate imbalance of the current capacities and to identify the potential of improved capacity dimensioning rules as introduced in Section 6.3.5, we use the simulation model from Section 6.3.6 to analyze system performance under various referral rates and dimensioning rules. For each scenario, we perform 15 simulation runs, each simulating 3000 weeks in the Pain Rehabilitation department, of which 500 weeks are taken as warm-up period. Applying Welch's

Table 6.2: Capacity requirements (in FTE) and bottleneck identification.

Discipline	*Current capacity (S)*	*Average capacity required per referral (A)*	*Capacity required for $\lambda = 17.7$ (D)*	*Load under $\lambda = 17.7$ (=D/S)*	*Maximum throughput (=S/A)*
RP	4.00	0.142	2.509	0.63	28.72
PT	7.38	0.859	15.212	2.06	8.59
OT	3.55	0.444	7.850	2.21	8.00
PS	4.53	0.398	7.048	1.56	11.38
SW	4.69	0.369	6.531	1.39	12.71
KI	2.58	0.151	2.664	1.03	17.14
Total	26.73	2.363	41.820	1.56	8.00

procedure [625], these settings provide a half-width of 5% for the 95% confidence intervals for the various performance indicators. The results are shown in Table 6.3.

We predict the performance of the system under its new design, which is different from current practice in several fundamental ways. First, the treatment plans have not yet been implemented in Het Roessingh. Second, no ad hoc flexibility in prioritizing patients will be allowed other than specified by type (i.e., intake, observation, group, individuals). Third, in the future it is only allowed to admit and schedule a patient when capacity is available with all disciplines for his/her entire treatment plan. Currently, only part of the treatment plan is scheduled at a time, and ad hoc adjustments are applied for patients who have been waiting too long, so that they receive a higher priority. Finally, indirect times will be included (reporting, administration) in the treatment time while in current practice it is possible to fill out doctors' schedules completely with treatments.

The third column of Table 6.3 represents the current capacity configuration. In Section 6.4.3, we have shown that multiple disciplines are overloaded when $\lambda = 17.7$, thus the queueing system under study is overloaded, implying that queue lengths and access times will grow rapidly without boundaries. Table 6.2 shows that with the current capacities, an upper bound on the maximum referral rate is 8.0 per week, dictated by the OT discipline. The inherent variability of demand and the interdependence between disciplines makes that that the referral rate must be even lower to keep access times and treatment delays stable. Experimenting with different referral rates revealed that $\lambda = 7.5$ would be manageable. With this load, we see in the second column of Table 6.3 that the utilization per discipline is varying between 26% and 93%, which is an indication of unbalanced capacity dimensioning. We note that in this case the waiting times for individual patients are highly fluctuating. Since scheduling individual treatments receives the lowest priority, this can be regarded as a sign that the system is operating close to its maximum capacity.

In columns 4 and 5 the results from aligning the number of FTEs per discipline to demand according to our proposed capacity dimensioning rule (6.1) with $\lambda = 7.5$ are shown, with the preset load 80% and 85%. We see that the average access times and total treatment times are comparable with the third column but that the load per discipline is better distributed and higher on average. Our proposed rule

Table 6.3: The results for multiple dimensioning rules (intake setting: $K = 4$, $N = 5$).

Referral rate (λ) →	*7.5*	*7.5*	*7.5*	*17.7*	*17.7*	*17.7*
Capacity dimensioning rule →	*current*	$\rho = 0.8$	$\rho = 0.85$	$\rho = 0.8$	$\rho = 0.85$	$\rho = 0.9$
Capacity (in FTE)						
RP	4.00	1.33	1.25	3.14	2.95	2.79
PT	7.38	8.06	7.58	19.02	17.90	16.90
OT	3.55	4.16	3.92	9.82	9.24	8.73
PS	4.53	3.73	3.51	8.81	8.29	7.83
SW	4.69	3.46	3.26	8.16	7.68	7.26
KI	2.58	1.41	1.33	3.33	3.13	2.96
Total	26.73	22.15	20.85	52.28	49.20	46.47
Load						
RP	0.27	0.80	0.85	0.80	0.85	0.90
PT	0.87	0.80	0.85	0.80	0.85	0.90
OT	0.94	0.80	0.85	0.80	0.85	0.90
PS	0.66	0.80	0.85	0.80	0.85	0.90
SW	0.59	0.80	0.85	0.80	0.85	0.90
KI	0.44	0.80	0.85	0.80	0.85	0.90
Overall	0.66	0.80	0.85	0.80	0.85	0.90
Performance: average (in weeks)						
Access time intake	1.4	1.4	1.5	1.1	1.1	1.2
Access time individual	2.2	0.2	0.7	0.0	0.1	0.7
Access time group	9.9	10.4	13.2	4.0	4.3	8.3
Access time observation	1.8	1.9	2.5	0.7	0.8	1.3
Prescribed treatment duration	21.3	21.3	21.3	21.3	21.3	21.2
Delay within treatment	4.1	3.5	4.8	1.3	1.4	3.3
Performance: 90^{th} percentile (in weeks)						
Access time intake	2.0	2.0	2.0	2.0	2.0	2.0
Access time individual	5.9	1.0	2.2	0.0	0.8	2.2
Access time group	22.6	23.8	28.1	9.4	10.0	14.9
Access time observation	5.0	5.0	6.3	2.0	2.0	3.3
Prescribed treatment duration	36.0	36.0	36.0	36.0	36.0	36.0
Delay within treatment	11.8	10.5	13.5	4.0	4.0	6.9

reduces the total number of FTE by 17% to 22% without decreasing system performance. Under a preset load per discipline of 90%, the system is overloaded. This is explained by the fact that disciplines are highly dependent: a discipline can often not start a treatment for one patient before the treatment of another patient is completed by another discipline. Therefore, some capacity loss is unavoidable. Here we also clearly observe the economies of scales effect: when $\lambda = 17.7$, the system can function under the preset load of 90% per discipline.

In the last three columns, the results are presented using our capacity dimensioning rule for $\lambda = 17.7$, with preset loads of 80%, 85% and 90%. We see that the loads per discipline are balanced and access times are small. Treatment delays decrease compared to the case with $\lambda = 7.5$, treatment discontinuity is thus reduced. A final observation is that to be able to fully implement treatment plans under a demand scenario of 17.7 referrals per week, a significant capacity increase is required.

6.5 Discussion

The results of this case study can be subdivided in two categories: basic insights and tools. The basic insights are the evaluation of the required number of intakes per week, and the quantitative characterization of the importance of multidisciplinary team meetings. In Section 6.4.1, we have shown that there is flexibility in how intake sessions can be organized, but the average number of available intake slots per week must be larger than an average number of weekly referrals by a visible margin, otherwise, the access times will inevitably increase beyond the Dutch national access time norm. To this end, we advise Het Roessingh to plan four intake sessions during each of which five patients can be seen. One could decide to plan more intake sessions but this will result in a decrease of the utilization of the planned intake sessions. In Section 6.4.2, we have demonstrated that cancelations of MTMs lead to unnecessary capacity losses. We conclude that the MTMs should have top priority, and the management must facilitate a high attendance of these meetings by the doctors and therapists.

We have developed tools to determine capacity requirements (Section 6.4.3) and to formulate capacity dimensioning rules (Section 6.4.4). These tools are tailored to the intended implementation of treatment plans. Indirect time is included in the treatment time to avoid clinician overload, and to ensure that MTMs are not canceled as a result of unavailability of the required documentation. Our methods allow to evaluate system design before actual changes are made. Therefore, they can assist the management in making optimal logistical decisions during the implementation of treatment plans.

To make patient flows more predictable, and to offer patients a consistent and reliable quality of care, more and more rehabilitation centers organize treatments according to treatment plans. We have shown that operations research methods, being able to quantitatively evaluate patients flows and prospectively asses the system's performance, are most suitable to support this organizational change. One of the main challenges of this study has been to map the treatment plans from the available data. We have described what type of data is essential for mapping, modeling and evaluating the patient flows. Our proposed methodology for the treatment plans mapping and system analysis, developed for Het Roessingh, is general which makes that it can be extended and applied in other rehabilitation centers.

This chapter illustrated the value of balanced capacity dimensioning in multidisciplinary treatment environments and Chapter 5 that of comprehensive treatment scheduling algorithms. An interesting direction for future research is to develop a methodology that combines the tools for balancing discipline capacities on the strategic level from this chapter with the tools for detailed treatment scheduling on the operational level from Chapter 5. Such integral decision making support, assists care facilities in realizing efficient organization of coordinated multidisciplinary treatments.

6.6 Appendix

6.6.1 Model for the intake process

This appendix belongs to Section 6.3.2. Let L_t denote the number of patients waiting for an intake at the beginning of week t, $t = 1, 2, \ldots$ Next, let A_t be the number of new patients (arrivals) in week t, and S_t the number of intakes (services) in week t. Then the queue length L_t satisfies a so-called Lindley's recursion:

$$L_{t+1} = L_t + A_t - S_t, \quad t = 1, 2, \ldots \tag{6.2}$$

Indeed, compared to the queue length at the beginning of week t, the number of arrivals must be added and the number of completed intakes must be subtracted in order to obtain the queue length at the beginning of week $t+1$. Furthermore, we note that S_t depends on L_t as follows:

$$S_t = \min\left\{N \cdot \left\lfloor \frac{L_t}{N} \right\rfloor, K \cdot N\right\}, \quad t = 1, 2, \ldots$$

This expression reflects the maximum number of sessions that can be filled, which has an upper limit of K, multiplied by the size of each session N. Assume that the number of arrivals in each week is independent of each other. Then $\{L_t, t \geq 1\}$ is a Markov process, of which the stationary probabilities, $\pi_l = \lim_{t\to\infty} P(L_t = l)$, $l \geq 0$, can be found.

We model A_t, $t = 1, 2, \ldots$, as independent random variables, which each have a Poisson distribution with mean λ. The assumption of Poisson distribution is common and most suitable if patients arrive independent of each other, as is the case in our study. We choose a truncation approximation, to approximate equation (6.2). Specifically, we assume that L_t cannot exceed some large but finite number l_{max} and numerically solve the balance equations of the resulting finite Markov chain. The obtained stationary probabilities provide a good approximation for the π_l's, and are then used to find the stationary probability distribution of the access time.

6.6.2 Model for multidisciplinary team meeting cancelations

This appendix belongs to Section 6.3.3. Let p denote the probability that an arbitrary MTM is canceled. For convenience of the mathematical presentation, we first present situation 2, and then situation 1.

Situation 2. If only the last meeting is important then extension of at least one period happens with probability p, and the probability of k six-week-extensions is $p^k(1-p)$, $k = 0, 1, \ldots$. This is a shifted geometric distribution with mean $p/(1-p)$.

Situation 1. If each canceled meeting leads to a six-weeks prolongation, then the treatment continues until all planned meetings have taken place. Since the treatment plan contains $x \geq 1$ meetings, the probability distribution of the number

of six-weeks-extensions is a sum of x random variables with a shifted geometric distribution with mean $p/(1-p)$.

6.6.3 Model for capacity requirements

This appendix belongs to Section 6.3.4. The calculations presented here, are performed for each discipline separately, therefore we omit a superscript d. Let $L^a_{w,j}$ denote the random appointment length, and $Z^a_{w,j}$ the mean amount of capacity (in hours) required per appointment of type a in week w. Then, clearly,

$$\mathbb{E}[Z^a_{w,j}] = n^a_{w,j} \cdot \mathbb{E}[L^a_{w,j}] \cdot p^a_{w,j},$$

and the mean capacity required per patient in week w of treatment plan j is

$$\mathbb{E}[Z_{w,j}] = \sum_a \mathbb{E}[Z^a_{w,j}].$$

Denote by $N_{w,j}$ the number of patients that, at a certain point in time, are in week w of a treatment plan j . Using the transition matrix and the mean number of arrivals per week, we are able to determine $\mathbb{E}[N_{w,j}]$ as follows. Consider all possible sequences of treatment plans $j_1 \to j_2 \to \cdots \to j_k \to j$ that lead to the treatment plan j. Assume also that all MTMs take place as planned, so there is no random prolongation of the treatment plans. Recall that λ is the average number of newly arrived patients per week. Then for each $w = 1, \ldots, l_j$, $j = 1, 2, \ldots, 18$, we have

$$\mathbb{E}[N_{w,j}] = \sum_{0 \to j_1 \to j_2 \to \cdots \to j_k \to j} \lambda \cdot Q_{(0,j_1)} Q_{(j_1,j_2)} Q_{(j_2,j_3)} \cdots Q_{(j_{k-1},j_k)} Q_{(j_k,j)}.$$

Denote by $C_{w,j}$ the amount of capacity required in week w for treatment plan j. Under the natural assumption that the numbers $N_{w,j}$ and $Z_{w,j}$ are independent, we find

$$\mathbb{E}[C_{w,j}] = \mathbb{E}[Z_{w,j}]\mathbb{E}[N_{w,j}].$$

The mean amount of capacity required to treat all patients, the *capacity requirements* for a discipline is given by

$$\mathbb{E}[C] = \sum_{j=1}^{18} \sum_{w=1}^{l_j} \mathbb{E}[C_{w,j}].$$

This approach can be extended to the case when MTMs are sometimes canceled leading to longer treatments, by allowing a random length of a treatment plan.

Part V

Integrally Shaping Inpatient Care Services

CHAPTER 7

Hourly Bed Census Predictions

7.1 Introduction

Inpatient care facilities provide care to hospitalized patients by offering a room, a bed and board [575]. Societal developments and budget constraints demand hospitals to on the one hand increase quality of care and on the other hand efficiency [494]. This entails a strong incentive to reconsider the design and operations of inpatient care services. In this chapter, we present an exact method to assist hospital management in adequately organizing their inpatient care services.

Effectively designing inpatient care services requires simultaneous consideration of several interrelated strategic and tactical planning issues (see Chapter 2). Given service mix and case mix decisions, hospital management has to decide on care unit partitioning (which care units are created and which patient groups are assigned to these units) and care unit size (the number of staffed beds per care unit). Since the inpatient care facility is a downstream department, the outflow of the operating theater and the emergency department, are main drivers behind its workload. Therefore, it is highly desirable to apply coordinated planning: considering the inpatient care facility in isolation yields suboptimal decision making [277, 591].

While smoothing patient inflow prevents large differences between peak and off-peak periods, and so realizes a more efficient use of resources [7, 277, 607], the authority of inpatient care facilities on their admission control is limited. Although the control on the inflow of patients from the emergency department is inherently very limited due to its nature, anticipation for emergency admissions is possible, by statistically predicting the arrival process of emergency patients that often follows a cyclic pattern [255]. Anticipation for elective surgical patients is possible as well, by taking the surgical schedule into account [7, 255, 277, 607]. Hospitals typically allocate operating room capacity through a Master Surgical Schedule (MSS), a (cyclic) block schedule that allocates operating time capacity among patient groups [210, 262, 589]. In this chapter, we address these various patient flows and take the necessity of integral decision making into account.

The challenge in decision making for inpatient care delivery is to guarantee care from appropriately skilled nurses and required equipment to patients with specific diagnoses, while making efficient use of scarce resources [282, 605]. Performance measures are required that reflect efficiency and quality of care to assess the quality

of logistical layout. Efficiency is often expressed in high bed occupancy, which is assumed to imply efficient use of staff and equipment [245, 500]. The drawback of high bed occupancy is that it may cause congestion, which manifests itself in two main consequences, both being a threat to the provided quality of care [247, 255]: (i) patients may have to be rejected for admission due to lack of bed capacity, so-called admission rejections, (ii) patients may (temporarily) be placed in less appropriate units, so-called misplacements [136, 281, 286]. Due to such misplacements, planning decisions regarding a specific care unit affects the operations of others [13, 124, 394]. Planning of inpatient care facilities should not only take into account the upstream departments, but also the interrelationship between care units.

Previous analytical studies have addressed partial resource capacity planning issues within the inpatient care chain, for example by dimensioning care units in isolation (e.g., [38, 245, 255]), balancing bed utilization across multiple units (e.g., [12, 124, 394]), or focusing on improving the MSS to balance inpatient care demand (e.g., [7, 39, 42, 589, 593]). More integral approaches can be found in simulation studies (e.g., [277, 281, 570, 590]). The advantage of such approaches is their flexibility, and therefore modeling power. However, the disadvantage is that the nature of such studies is typically context specific, which limits the generalizability of application and findings.

We present a generic exact analytical approach to achieve the required integral and coordinated resource capacity planning decision-making for inpatient care services. The method builds upon the approach presented in [593], which determines the workload placed on hospital departments by describing demand for elective inpatient care beds on a daily level as a function of the MSS. Based on a cyclic arrival pattern of emergency patients and an MSS block schedule of surgical patients, we derive demand predictions on an hourly level for several inpatient care units simultaneously for both acute and elective patients. (The method is also applicable for departments catering for non-surgical elective patients, as these can be incorporated in our model via fictitious OR blocks). This hourly level of detail is required to adequately incorporate the time-dependent behavior of the inpatient care process. Based on overflow rules we translate the demand predictions to bed census predictions, since demand and census may differ due to rejections and misplacements. The combination of the hourly level perspective and the bed census conversion enables us to derive several performance measures, along which the effectiveness of different logistical configurations can be assessed. In addition, what-if questions can be addressed considering the impact of operational interventions such as shortening length of stay or changing the times of admissions and discharges.

During the upcoming years the presented method will be applied in the Academic Medical Center (AMC) Amsterdam in supporting the intended complete redesign of the inpatient care facility. As part of the total redesign, in the case study we present here we restrict ourselves to a set of interrelated (with respect to capacity planning) specialties: traumatology, orthopedics, plastic surgery, urology, vascular surgery, and general surgery. By means of this case study we illustrate the practical potential of our analytical approach for logistical redesign of inpatient care services.

This chapter is organized as follows. First, Section 7.2 introduces the case study at the AMC. In Section 7.3, we describe the model consisting of demand predictions, bed census predictions and performance measures. Section 7.4 presents the numerical results. The chapter closed with a discussion of our findings and opportunities for further research in Section 7.5.

7.2 Background: case study

The case study entails the university hospital AMC, which has 20 operating rooms, and 30 inpatient departments with in total 1000 beds. Due to both economic and medical developments, the AMC is forced to reorganize the operations of the inpatient services during the upcoming years. On the basis of an example for six surgical specialties, we the potential of the presented method to direct these reorganizations.

The following specialties are taken into account: traumatology (TRA), orthopedics (ORT), plastic surgery (PLA), urology (URO) vascular surgery (VAS), and general surgery (GEN). In the present setting, the patients of the mentioned specialties are admitted in four different inpatient care departments. Care unit A houses GEN and URO, unit B VAS and PLA, unit C TRA, and unit D ORT. The physical building is such that units A and B are physically adjacent (Floor I), so are units C and D (Floor II). For these specialties, we have historical data available over 2009–2010 on 3498 (5025) elective (acute) admissions, with an average length of stay (LOS) of 4.85 days (see Table 7.1). Currently, no cyclical MSS is applied. Each time, roughly six weeks in advance the MSS is determined for a period of four weeks.

The capacities of units A, B, C, and D are 32, 24, 24, and 24 beds, respectively. However, it often happens that not all beds are available, due to personnel shortages. The utilizations over 2009–2010 were 53.2%, 55.6%, 54.4%, and 60.6% (which includes some patients of other than the given specialties that were placed in these care units). These utilizations reflect administrative bed census, which means the percentage of time that a patient physically occupies a bed, or keeps it reserved during the time the patient is at the operating theater or at the intensive care department. Unfortunately, no confident data was available on rejections and misplacements.

Table 7.1: Overview historical data 2009-2010.

Specialty	*Acronym*	*Care unit*	*Elective admissions*	*Acute admissions*	*Average LOS (in days)*	*Load (# patients)*
General surgery	GEN	A	611	901	3.31	6.88
Urology	URO	A	818	1157	3.68	9.99
Vascular surgery	VAS	B	257	634	8.30	10.16
Plastic surgery	PLA	B	639	288	2.29	2.91
Traumatology	TRA	C	337	1200	5.88	12.41
Orthopedics	ORT	D	836	845	6.23	14.38

7.3 Methods

In this section, the model is described that predicts the workload at several care units of an inpatient care facility on a time scale of hours, due to patients originating from the operating theater and emergency department. The basis for the operating room outflow prediction is the MSS. The basis for the emergency department outflow prediction is a cyclic random arrival process which we define as the Acute Admission Cycle (AAC). Schematically, the approach is as follows. First, the impact of the MSS and the AAC are separately determined and then combined to obtain the overall steady state impact of the repeating cycles. Second, the obtained demand distributions are translated to bed census distributions. Finally, performance measures are formulated based on the demand and census distributions.

The operation of the inpatient care facility is as follows. Each day is divided in time intervals, which in principle can be regarded as hours (but could also resemble for example two- or four-hour time intervals). Patient admissions are assumed to take place independently at the start of a time interval. Elective patients are admitted to a care unit either on the day before or on the day of surgery. For acute patients we assume a cyclic (e.g., weekly) non-homogeneous Poisson arrival process corresponding to the unpredictable nature of emergency arrivals. Discharges take place independently at the end of a time interval. For elective patients we assume the length of stay to depend only on the type of patient and to be independent of the day of admission and the day of discharge. For acute patients the length of stay and time of discharge are dependent on the day and time of arrival, in particular to account for possible disruptions in diagnostics and treatment during nights and weekends.

For the demand predictions, for both elective and acute patients three steps are performed. First, the impact of a single patient type in a single cycle (MSS or AAC) is determined, by which in the second step the impact of all patient types within a single cycle can be calculated. Then, since the MSS and AAC are cyclical, the predictions from the second step are overlapped to find the overall steady state impact of the repeating cycles. The workload predictions for elective and acute patients are combined to find the probability distributions of the number of recovering patients at the inpatient care facility on each unique day in the cycle which we denote as the Inpatient Facility Cycle (IFC). The length of the IFC is the least common multiple of the lengths of the MSS and the AAC.

Patient admission requests may have to be rejected due to a shortage of beds, or patients may (temporarily) be placed in less appropriate units. As a consequence, demand predictions and bed census predictions do not coincide. Therefore, an additional step is required to translate the demand distributions into census distributions. This translation is performed by assuming that after a misplacement the patient is transferred to his preferred care unit when a bed becomes available, where we assume a fixed patient-to-ward allocation policy, which prescribes the prioritization of such transfers.

7.3.1 Demand predictions for elective patients

Model input. The demand predictions for elective patients will be based on the following input parameters.

Time. An MSS is a repeating blueprint for the surgical schedule of S days. Each day is divided in T time intervals. Therefore, we have time points $t = 0, \ldots, T$, in which $t = T$ corresponds to $t = 0$ of the next day. For each single patient, day n counts the number of days before or after surgery, i.e., $n = 0$ indicates the day of surgery.

MSS utilization. For each day $s \in \{1, \ldots, S\}$, a (sub)specialty j can be assigned to an available operating room i, $i \in \{1, \ldots, I\}$. The OR block at operating room i on day s is denoted by $b_{i,s}$, and is possibly divided in a morning block $b^m_{i,s}$ and an afternoon block $b^a_{i,s}$, if an OR day is shared. The discrete distributions c^j represent how specialty j utilizes an OR block, i.e., $c^j(k)$ is the probability of k surgeries performed in one block, $k \in \{0, 1, ..., C^j\}$. If an OR block is divided in a morning OR block and an afternoon OR block, c^j_M and c^j_A represent the utilization probability distributions respectively. We do not include shared OR blocks explicitly in our formulation, since these can be modeled as two separate (fictitious) operating rooms.

Admissions. With probability e^j_n, $n \in \{-1, 0\}$, a patient of type j is admitted on day n. Given that a patient is admitted on day n, the time of admission is described by the probability distribution $w^j_{n,t}$. We assume that a patient who is admitted on the day of surgery is always admitted before or at time ϑ_j; therefore, we have $w^j_{0,t} = 0$ for $t = \vartheta_j + 1, \ldots, T - 1$.

Discharges. $P^j(n)$ is the probability that a type j patient stays n days after surgery, $n \in \{0, \ldots, L^j\}$. Given that a patient is discharged on day n, the probability of being discharged in time interval $[t, t + 1)$ is given by $m^j_{n,t}$. We assume that a patient who is discharged on the day of surgery is discharged after time ϑ_j, i.e., $m^j_{0,t} = 0$ for $t = 0, \ldots, \vartheta_j$.

Single surgery block. In this first step we consider a single specialty j operating in a single OR block. We compute the probability $h^j_{n,t}(x)$ that n days after carrying out a block of specialty j, at time t, x patients of the block are still in recovery. Note that admissions can take place during day $n = -1$ and during day $n = 0$ until time $t = \vartheta_j$. Discharges can take place during day $n = 0$ from time $t = \vartheta_j + 1$ and during days $n = 1, \ldots, L^j$. Therefore, we calculate $h^j_{n,t}(x)$ as follows:

$$h^j_{n,t}(x) = \begin{cases} a^j_{n,t}(x) & \text{, if } n = -1 \text{ and } n = 0, t \leq \vartheta_j, \\ d^j_{n,t}(x) & \text{, if } n = 0, t > \vartheta_j \text{ and } n = 1, \ldots, L^j, \end{cases}$$

where $a^j_{n,t}(x)$ represents the probability that x patients are admitted until time t on day n, and $d^j_{n,t}(x)$ is the probability that x patients are still in recovery at time t on day n. The derivations of $a^j_{n,t}$ and $d^j_{n,t}$ are presented in Appendix 7.6.1.

Single MSS cycle. Now, we consider a single MSS in isolation. From the distributions $h^j_{n,t}$, we can determine the distributions $H_{m,t}$, the discrete distributions for the total number of recovering patients at time t on day m, $m \in \{0,1,2,\ldots,S,S+1,S+2,\ldots\}$, resulting from a single MSS cycle (see Appendix 7.6.1).

Steady state. In this step, the complete impact of the repeating MSS is considered. The distributions $H_{m,t}$ are used to determine the distributions $H^{SS}_{s,t}$, the steady state probability distributions of the number of recovering patients at time t on day s of the cycle, $s \in \{1,\ldots,S\}$ (see Appendix 7.6.1).

7.3.2 Demand predictions for acute patients

Model input. The demand predictions for acute patients will be based on the following input parameters.

Time. The AAC is the repeating cyclic arrival pattern of acute patients with a length of R days. For each single patient, day n counts the number of days after arrival.

Admissions. An acute patient type is characterized by patient group p, $p = 1,\ldots,P$, arrival day r and arrival time θ, which is for notational convenience denoted by type $j = (p,r,\theta)$. The Poisson arrival process of patient type j has arrival rate λ^j.

Discharges. $P^j(n)$ denotes the probability that a type j patient stays n days, $n \in \{0,\ldots,L^j\}$. Given that a patient is discharged at day n, the probability of being discharged in time interval $[t,t+1)$ is given by $\tilde{m}^j_{n,t}$. By definition, $\tilde{m}^j_{0,t} = 0$ for $t \leq \theta$.

Single patient type. In this first step we consider a single patient type j. We compute the probability $g^j_{n,t}(x)$ that on day n at time t, x patients are still in recovery. Admissions can take place during time interval $[\theta,\theta+1)$ on day $n = 0$ and discharges during day $n = 0$ after time θ and during days $n = 1,\ldots,L^j$. Therefore, we calculate $g^j_{n,t}(x)$ as follows:

$$g^j_{n,t}(x) = \begin{cases} \tilde{a}^j_t(x) & \text{, if } n = 0, t = \theta, \\ \tilde{d}^j_{n,t}(x) & \text{, if } n = 0, t > \theta \text{ and } n = 1,\ldots,L^j, \end{cases}$$

where $\tilde{a}^j_t(x)$ represents the probability that x patients are admitted in time interval $[t,t+1)$ on day $n = 0$, and $\tilde{d}^j_{n,t}(x)$ is the probability that x patients are still in recovery at time t on day n. In Appendix 7.6.2, we present the derivations of $\tilde{a}^j_t$ and $\tilde{d}^j_{n,t}$.

Single cycle. Now, we consider a single AAC in isolation. From the distributions $g^j_{n,t}(x)$, we can determine the distributions $G_{w,t}$, the distributions for the total number of recovering patients at time t on day w, $w \in \{1,\ldots,R,R+1,R+2,\ldots\}$, resulting from a single AAC (see Appendix 7.6.2).

Steady state. In this step, the complete impact of the repeating AAC is considered. The distributions $G_{w,t}$ are used to determine the distributions $G^{SS}_{r,t}$, the steady state probability distributions of the number of recovering patients at time t on day r of the cycle, $r \in \{1,\dots,R\}$ (see Appendix 7.6.2).

7.3.3 Demand predictions per care unit

To determine the complete demand distribution of both elective and acute patients, we need to combine the steady state distributions $H^{SS}_{s,t}$ and $G^{SS}_{r,t}$. In general, the MSS cycle and AAC are not equal in length, i.e., $S \neq R$. This has to be taken into account when combining the two steady state distributions. Therefore, we define the new IFC length $Q = LCM(S,R)$, where the function LCM stands for *least common multiple*. Let $Z_{q,t}$ be the probability distribution of the total number of patients recovering at time t on day q during a time cycle of length Q:

$$Z_{q,t} = H^{SS}_{q \bmod S + S\cdot \mathbb{1}_{(q \bmod S=0)},t} \otimes G^{SS}_{q \bmod R + R\cdot \mathbb{1}_{(q \bmod R=0)},t},$$

where $\otimes$ denotes the discrete convolution function. Let W^k be the set of specialties j whose operated patients are (preferably) admitted to unit k, $k \in \{1,\dots,K\}$, and V^k the set of acute patient types j that are (preferably) admitted to unit k. Then, the demand distribution for unit k, $Z^k_{q,t}$, can be calculated by exclusively considering the patients in W^k in equation (7.7) and V^k in equation (7.8).

7.3.4 Bed census predictions

We translate the demand distributions $Z^k_{q,t}$ into bed census distributions $\hat{Z}_{q,t}$, $k = 1,\dots,K$, the distributions of the number of patients present in each unit k at time t on day q. To this end, we require an allocation policy ϕ that uniquely specifies from a demand vector $\mathbf{x} = (x_1,\dots,x_K)$ a bed census vector $\hat{\mathbf{x}} = (\hat{x}_1,\dots,\hat{x}_K)$, in which x_k and $\hat{x}_k$ denote the demand for unit k and the bed census at unit k, respectively. Let $\phi(\cdot)$ be the function that executes allocation policy ϕ. Let $\hat{Z}^k_{q,t}$ denote the marginal distribution of the census at unit k given by distribution $\hat{Z}_{q,t}$. With M^k the capacity of unit k in number of beds, we obtain

$$\hat{Z}_{q,t}(\hat{\mathbf{x}}) = \left(\hat{Z}^1_{q,t}(\hat{x}_1),\dots,\hat{Z}^K_{q,t}(\hat{x}_K)\right) = \sum_{\{\mathbf{x}|\hat{\mathbf{x}}=\phi(\mathbf{x})\}} \left\{\prod_{k=1}^{K} Z^k_{q,t}(x_k)\right\}. \tag{7.1}$$

We do not impose restrictions on the allocation policy ϕ other than specifying a unique relation between demand $\mathbf{x}$ and census configuration $\hat{\mathbf{x}}$. Recall that the underlying assumption is that a patient is transferred to his preferred unit when a bed becomes available. The policy ϕ also reflects the priority rules that are applied for such transfers. As an illustration, we present an example for an inpatient care facility

with two care units of capacity M^1 and M^2 respectively:

$$\phi(\mathbf{x}) = \begin{cases} (x_1, x_2) & \text{, if } x_1 \leq M_1, x_2 \leq M_2, \\ (M_1, \min\{x_2 + (x_1 - M_1), M_2\}) & \text{, if } x_1 > M_1, x_2 \leq M_2, \\ (\min\{x_1 + (x_2 - M_2), M_1\}, M_2) & \text{, if } x_1 \leq M_1, x_2 > M_2, \\ (M_1, M_2) & \text{, if } x_1 > M_1, x_2 > M_2. \end{cases} \tag{7.2}$$

Under this policy patients are assigned to their bed of preference if available, and are otherwise misplaced to the other unit if beds are available there.

7.3.5 Performance indicators

Based on the demand distributions $Z^k_{q,t}$ and the census distributions $\hat{Z}^k_{q,t}$ we are able to formulate a variety of performance indicators. We present a selection of such performance indicators, which will be used in the next section to evaluate the impact of different scenarios and interventions.

Demand percentiles. Let $D^k_{q,t}(\alpha)$ be the α-th demand percentile at time t on day q:

$$D^k_{q,t}(\alpha) = \min_x \Big\{ \sum_{i=0}^{x} Z^k_{q,t}(i) \geq \alpha \Big\}.$$

(Off-)Peak demand. Reducing peaks and drops in demand will balance bed occupancy and therefore allows more efficient use of available staff and beds. Define $\overline{P}^k_q(\alpha)$ ($\underline{P}^k_q(\alpha)$) and $\overline{P}^k(\alpha)$ ($\underline{P}^k(\alpha)$) to be the maximum (minimum) α-th demand percentile per day and over the complete cycle respectively:

$$\overline{P}^k_q(\alpha) = \max_t \{D^k_{q,t}(\alpha)\}, \quad \overline{P}^k(\alpha) = \max_q \{\overline{P}^k_q(\alpha)\},$$
$$\underline{P}^k_q(\alpha) = \min_t \{D^k_{q,t}(\alpha)\}, \quad \underline{P}^k(\alpha) = \min_q \{\underline{P}^k_q(\alpha)\}.$$

Admission rate. Patient admissions may increase the nursing workload. Let $\Lambda^k_{q,t}$ be the distribution of the number of arriving patients during time interval $[t, t+1)$ on day q who are preferably admitted to care unit k. To obtain $\Lambda^k_{q,t}$, we first determine $\bar{a}^j_{n,t}$, the distribution of the number of elective type j arrivals during time interval $[t, t+1)$ on day n ($n \in \{-1, 0\}$):

$$\bar{a}^j_{n,t}(x) = \sum_{y=0}^{C^j} c^j(y) \bar{a}^j_{n,t}(x|y), \text{with}$$
$$\bar{a}^j_{n,t}(x|y) = \binom{y}{x} (e^j_n w^j_{n,t})^x (1 - e^j_n w^j_{n,t})^{y-x}.$$

$\Lambda^k_{q,t}$ is then determined by taking the discrete convolution over all relevant arrival distributions of both elective and acute patient types:

$$\Lambda^k_{q,t} = \Big\{ \bigotimes_{i=1}^{I} \Big\{ \bigotimes_{j \in W^k : j \in b_{i,s'}} \bar{a}^j_{-1,t} \Big\} \otimes \Big\{ \bigotimes_{j \in W^k : j \in b_{i,s''}} \bar{a}^j_{0,t} \Big\} \Big\} \otimes \Big\{ \bigotimes_{j \in V^k : r = r'} \tilde{a}^j_t \Big\}. \tag{7.3}$$

where $s' = 1 + q \bmod S$, $s'' = q \bmod S + S \cdot \mathbb{1}_{(q \bmod S = 0)}$, $r' = q \bmod R + R \cdot \mathbb{1}_{(q \bmod R = 0)}$, and $\bigotimes_{x \in \mathcal{X}} f_x$ denotes the discrete convolution over the probability distributions $f_x, x \in \mathcal{X}$. The first term in the right-hand side of (7.3) represents the elective patients who claim a bed at unit k ($j \in W^k$), who are operated in any OR and who are admitted on the day $s' - 1$ before surgery or on the day s'' of surgery. The second term in the right-hand side of (7.3) represents the acute patients who claim a bed at unit k ($j \in V^k$) and who arrive on the corresponding day r' in the AAC.

Average bed occupancy. Let $\rho^k_{q,t}$, ρ^k_q, ρ^k be the average number of beds occupied at care unit k respectively at time t on day q, on day q, and over the complete cycle:

$$\rho^k_{q,t} = \frac{1}{M^k} \sum_{x=0}^{M^k} x \cdot \hat{Z}^k_{q,t}(x), \qquad \rho^k_q = \frac{1}{T} \sum_{t=0}^{T-1} \rho^k_{q,t}, \qquad \rho^k = \frac{1}{Q} \sum_{q=1}^{Q} \rho^k_q.$$

Rejection probability. Let $R^{\phi,k}$ denote the probability that under allocation policy ϕ an admission request of an arriving patient for unit k has to be rejected, because all beds at unit k are already occupied and none of the alternative beds (prescribed by ϕ) are available. To determine $R^{\phi,k}$, we first determine $R^{\phi,k}_{q,t}$: the probability of such an admission rejection at time t on day q. $R^{\phi,k}$ is then calculated as follows:

$$R^{\phi,k} = \frac{1}{\sum_{q,t} E[\Lambda^k_{q,t}]} \sum_{q,t} E[\Lambda^k_{q,t}] R^{\phi,k}_{q,t}.$$

Let n indicate the number of arriving patients who are preferably admitted to unit k, and $\mathbf{x} = (x_1, \ldots, x_K)$ the demand for each unit (in which these arrivals are already incorporated). Introduce $\mathcal{R}^{\phi,k}(\mathbf{x}, n)$, the number of rejected patients under allocation policy ϕ of the n arriving patients to unit k, and $Z^k_{q,t}(x_k|n)$ the probability that at time t on day q in total x_k patients demand a bed at unit k and n of them have just arrived. Then, $R^{\phi,k}_{q,t}$ is calculated by:

$$\begin{aligned} R^{\phi,k}_{q,t} &= \frac{E[\#\text{ rejections at unit } k \text{ on time } (q,t)]}{E[\#\text{ arrivals to unit } k \text{ on time } (q,t)]} \\ &= \frac{1}{E[\Lambda^k_{q,t}]} \sum_{\mathbf{x}} \prod_{\ell \neq k} Z^\ell_{q,t}(x_\ell) \sum_n \mathcal{R}^{\phi,k}(\mathbf{x}, n) \Lambda^k_{q,t}(n) Z^k_{q,t}(x_k|n). \end{aligned} \tag{7.4}$$

The derivation of $Z^k_{q,t}(x_k|n)$ is presented in Appendix 7.6.3. $\mathcal{R}^{\phi,k}(\mathbf{x}, n)$ is uniquely determined by allocation policy ϕ. For example, for the case with $K = 2$ presented in (7.2), we have for unit $k = 1$:

$$\mathcal{R}^{\phi,1}(\mathbf{x}, n) = \begin{cases} \min\{n, x_1 - M_1\}, & \text{if } x_1 \geq M_1, x_2 \geq M_2, \\ \max\{0, (x_1 - M_1) - (M_2 - x_2)\}, & \\ & \text{if } x_1 \geq M_1, x_2 < M_2, n \geq (x_1 - M_1), \\ n - [\min\{n, (M_2 - x_2 - [x_1 - M_1 - n])\}]^+, & \\ & \text{if } x_1 \geq M_1, x_2 < M_2, n < (x_1 - M_1), \\ 0, & \text{otherwise.} \end{cases}$$

Here, the first case reflects the situation in which all beds at care unit 2 are occupied so that all arriving patients who do not fit in unit 1 have to be rejected. The second and third case reflect the situation that (some of) the arriving patients can be misplaced to unit 2 so that only a part of the arriving patients have to be rejected. In the second case, the $(x_1 - M_1)$ patients that do not fit at unit 1 are all arriving patients. In the third case, some of the $(x_1 - M_1)$ patients were already present so that not all $(M_2 - x_2)$ beds at unit 2 can be used to misplace arriving patients.

Misplacement probability. Let $M^{\phi,k}$ denote the probability that under allocation policy ϕ a patient who is preferably admitted to care unit k is admitted to another unit. The derivation of $M^{\phi,k}$ is equivalent to that of $R^{\phi,k}$. In (7.4), $\mathcal{R}^{\phi,k}(\mathbf{x}, n)$ has to be replaced by $\mathcal{M}^{\phi,k}(\mathbf{x}, n)$, which gives the number of misplaced patients under allocation policy ϕ of the n arriving patients to unit k and which is again uniquely determined by ϕ. Observe that for the two unit example presented in (7.2), we have:

$$\mathcal{M}^{\phi,1}(\mathbf{x}, n) = \begin{cases} \min\{x_1 - M_1, M_2 - x_2\}, & \\ & \text{if } x_1 > M_1, x_2 < M_2, n \geq (x_1 - M_1), \\ \max\left\{0, \min\{n, (M_2 - x_2 - [x_1 - M_1 - n])\}\right\}, & \\ & \text{if } x_1 > M_1, x_2 < M_2, n < (x_1 - M_1), \\ 0, & \text{otherwise.} \end{cases}$$

Productivity. Let $\mathcal{K}$ be a set of cooperating care units, i.e., units that mutually allow misplacements. Let $\mathcal{P}^{\mathcal{K}}$ reflect the productivity of the available capacity at care units $k \in \mathcal{K}$, defined as the number of patients that is treated per bed per day:

$$\mathcal{P}^{\mathcal{K}} = \frac{365}{Q} \frac{1}{\sum_{k \in \mathcal{K}} M^k} \sum_{k \in \mathcal{K}} \sum_{q,t} (1 - R^{\phi,k}_{q,t}) E[\Lambda^k_{q,t}]. \tag{7.5}$$

Remark 7.1 (Approximation)**.** Observe that the calculations of misplacements and rejections are an abstract approximation of complex reality. In our model, we count each time interval how many of the arriving patients have to be misplaced or rejected. Since we do not remove rejected patients from the demand distribution, it is likely that we overestimate the rejection and misplacement probabilities. However, also in reality strict rejections are often avoided: by postponing elective admissions, predischarging another patient, or letting acute patients wait at the emergency department. These are all undesired degradations of provided quality of care. Therefore, our method provides a secure way of organizing inpatient care services. It is applicable to evaluate performance for care unit capacities that give low rejection probabilities, thus when high service levels are desired, which is typically the case in healthcare.

Remark 7.2 (Numerical evaluation)**.** Recall that to compute all performance measures formulated above it is only required to specify the input parameters that were specified under the headers 'model input' for the elective and the acute patients.

7.4 Numerical results

In this section, we illustrate the practical potential of our analytical approach for logistical redesign of inpatient care services, by means of the case study introduced in Section 7.2.

7.4.1 Validation

We have estimated the input parameters for our model based on historical data of 2009–2010 from the hospital electronic databases. The event logs of the operating room and inpatient care databases had to be matched. Since the data contained many errors, extensive cleaning was required. Patients of other specialities who stayed at departments A–D have been deleted. Since no cyclical MSS was applied in practice, we set the MSS length on two years, following the surgery blocks as occurred in practice during 2009–2010. Elective surgery blocks were only executed on weekdays. For the elective patient types, the distributions for the number of surgeries and for the admission/discharge processes are estimated per specialty. We set the length of the AAC on one week. For the acute patients, the discharge distributions are estimated per specialty, and to have enough measurements, via the following clustering: admission time intervals 0–8, 8–18, and 18–24. Furthermore, for all patient types the discharge distributions during a day are assumed to be equal for the days $n \geq 2$.

As an illustration, Figure 7.1 displays the model results for demand distributions $Z_{q,t}^{k}$ for care unit A on Wednesdays against the historical data. The results are similar for the other days and the other care units. Slight differences can be observed for (1) the elective patients on Sunday afternoon, since in practice Sunday-admission times differ from weekdays, where we assume the same admission time distributions for all days, and (2) the elective patients on Friday afternoon, since in practice more patients are discharged just before the weekend, where we assume the length-of-stay distributions to be independent of the day of surgery. We conclude that our model is a valid representation of the AMC practice.

7.4.2 Analysis

We consider several interventions which potentially improve the efficiency of the inpatient care service operations. For the interventions that are based on the current MSS, we run the model for the estimated two-year MSS, and we calculate the performance measures only over the second year, to account for warm-up effects. To asses the effects of the interventions, we first evaluate the performance of a base case scenario, the situation that most closely resembles current practice. The base case takes the current capacities, misplacements take place between care units A and B (floor I), and between units C and D (floor II). We assume that the available beds are always open, so no ad-hoc bed closings are allowed. Note that the calculated rejection and misplacement percentages are therefore most likely an underestima-

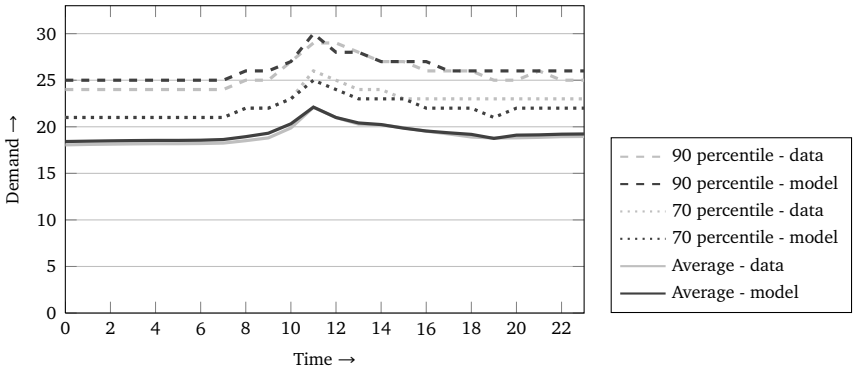


Figure 7.1: Validation of the model results against historical data (Wednesday, care unit A).

tion of current practice (of which no reliable data is available). The productivity measure is calculated per floor, since the misplacement policy implies that capacity is 'shared' per floor. The following interventions are considered, of which the results are displayed in Tables 7.2-7.4:

(1) Rationalize bed requirements. The current numbers of beds are a result of historical development. Given particular service requirements, which are to be specified by the hospital, we determine whether the number of beds can be reduced to achieve a higher bed utilization while a certain quality level is guaranteed. We consider rejection probabilities not exceeding 5%, 2.5%, and 1%. Often, there are different bed configurations with the same total number of beds per floor, satisfying a given maximum rejection probability. Per floor, from the available configurations the one is chosen that gives the lowest maximum misplacement probability.

It can be seen that a significant reduction in the number of beds is possible. However, the overall bed utilizations are still modest, because demand drops during weekend days when no elective surgeries take place. In addition, there is a correlation between moments of higher census and moments that patients arrive, which leads to higher rejection probabilities compared to for instance a stationary Poisson arrival process. The hospital recognizes that simultaneously prohibiting bed closings on an ad-hoc basis and downsizing the total number of beds is more effective in realizing a consistent quality-of-service level, whilst it is also more efficient (reflected by the clear increase in the productivity measure, i.e., the number of patients that can be treated per bed per day).

(2) No misplacements. For the purpose of insight, in this intervention we explore what would happen if no misplacements were allowed. By banning misplacements, we demonstrate the benefits of capacity pooling when overflow between units is allowed. These benefits are due to the so-called portfolio effect which induces that the relative variability in demand is reduced by economies of scale.

It can be concluded that in our case units in the order of size 20–30 beds are too small to operate efficiently in isolation.

Table 7.2: The numerical results for the base case, intervention 1, and intervention 2 (with the productivity-Δ% relative to the base case).

Intervention	*Unit*	*Capacity* (# beds)	*Rejection* (%)	*Misplace* (%)	*Utilization* (%)	*Floor*	*Capacity* (# beds)	*Productivity* (eq.(7.5))	(Δ%)
Base case	A	32	0.14	1.85	56.9	AB	56	50.0	-
	B	24	0.08	1.22	56.5				
	C	24	0.03	0.45	55.6	CD	48	35.1	-
	D	24	0.10	3.68	61.5				
1. Rationalize bed requirements									
Rejection < 5%	A	27	4.92	6.07	67.7	AB	45	59.3	+18.6
	B	18	4.59	14.35	74.3				
	C	18	3.42	8.90	74.0	CD	38	42.5	+21.1
	D	20	4.92	11.72	73.3				
Rejection < 2.5%	A	28	2.31	5.86	65.0	AB	48	57.2	+14.4
	B	20	1.67	7.30	67.7				
	C	18	2.02	10.30	73.3	CD	40	41.3	+17.5
	D	22	2.27	6.14	67.5				
Rejection < 1%	A	29	0.94	5.00	62.6	AB	51	54.5	+9.1
	B	22	0.52	3.15	61.8				
	C	20	0.54	4.39	66.5	CD	43	39.0	+11.0
	D	23	0.79	4.93	64.3				
2. No misplacements									
Rejection < 5%	A	30	4.22	-	60.5	AB	52	51.7	+3.5
	B	22	3.67	-	61.5				
	C	20	4.93	-	66.1	CD	44	36.7	+4.4
	D	24	3.78	-	61.5				
Rejection < 2.5%	A	32	2.00	-	56.8	AB	55	49.9	-0.2
	B	23	2.22	-	58.9				
	C	22	1.67	-	60.3	CD	47	35.2	+0.1
	D	25	2.42	-	59.1				
Rejection < 1%	A	34	0.86	-	53.5	AB	59	47.1	-5.7
	B	25	0.73	-	54.2				
	C	23	0.91	-	57.8	CD	50	33.4	-4.8
	D	27	0.90	-	54.8				

(3) Change operational process. First, hospital management proposes to admit all elective patients on the day of surgery, since admitting patients the day before surgery is often induced by logistical reasons and not by medical necessity. Second, to reduce census peaks during the middle of the day, management proposes to aim for discharges to happen before noon. To predict the potential impact of these changes in the operational process, we adjust the admission distributions of elective patients, so that admissions on the day before surgery are postponed to time $t = 8$ on the day of surgery (which impacts 81.9% of the elective patients), and we adjust the discharge distributions of days $n \geq 1$, so that discharges later than time $t = 11$ are moved forward to $t = 11$ (which impacts 51.8% of the total patient population).

Table 7.3: The numerical results for interventions 2, 3 and 4 (with the productivity-Δ% relative to the base case).

Intervention	*Unit*	*Capacity* (# beds)	*Rejection* (%)	*Misplace* (%)	*Utilization* (%)	*Floor*	*Capacity* (# beds)	*Productivity* (eq.(7.5))	(Δ%)
3. Change operational process									
Rejection < 5%	A	24	4.51	9.24	66.4	AB	43	62.5	+25.2
	B	19	3.03	6.53	66.1				
	C	17	3.65	11.21	74.3	CD	37	43.6	+24.2
	D	20	5.00	9.12	69.7				
Rejection < 2.5%	A	26	2.31	5.22	61.7	AB	45	60.9	+21.8
	B	19	2.03	7.54	65.7				
	C	17	2.11	12.74	73.8	CD	39	42.3	+20.5
	D	22	2.28	4.62	64.0				
Rejection < 1%	A	27	0.94	4.44	59.3	AB	48	57.9	+15.8
	B	21	0.64	3.26	59.7				
	C	19	0.58	5.59	66.8	CD	42	39.9	+13.6
	D	23	0.83	3.78	60.7				
4. Balance MSS									
Rejection < 5%	A	25	4.85	8.43	74.5	AB	44	62.5	+25.0
	B	19	3.93	8.73	74.4				
	C	18	3.24	8.84	74.6	CD	38	43.5	+23.7
	D	20	3.99	10.03	75.6				
Rejection < 2.5%	A	27	2.25	4.29	69.5	AB	46	61.1	+22.3
	B	19	2.41	10.25	73.9				
	C	19	1.46	6.21	70.8	CD	40	42.1	+19.9
	D	21	1.86	7.50	72.2				
Rejection < 1%	A	28	0.83	3.57	66.7	AB	49	58.3	+16.6
	B	21	0.66	4.32	67.4				
	C	20	0.60	4.05	67.3	CD	42	40.5	+15.3
	D	22	0.79	5.21	69.0				
5. Combination (1), (3), and (4)									
Rejection < 5%	A	23	4.92	9.17	70.9	AB	42	65.5	+31.1
	B	19	3.47	5.56	68.9				
	C	17	3.77	11.04	74.9	CD	37	44.5	+26.5
	D	20	4.21	7.34	71.7				
Rejection < 2.5%	A	25	2.28	4.72	65.7	AB	44	64.0	+28.0
	B	19	2.18	6.85	68.4				
	C	18	1.74	7.87	71.0	CD	39	43.1	+22.7
	D	21	2.02	5.54	68.2				
Rejection < 1%	A	26	0.82	3.90	63.1	AB	47	60.8	+21.7
	B	21	0.57	2.75	62.2				
	C	19	0.74	5.21	67.5	CD	41	41.4	+18.0
	D	22	0.89	3.87	65.1				

Compared to intervention 1 the number of beds can be further decreased. Also, the results indicate that hospitals should not only focus on achieving high bed utilizations: although somewhat lower utilization is achieved, productivity is significantly increased.

(4) Balance MSS. The realized MSS created artificial demand variability. This intervention estimates the potential of a cyclical MSS that is designed with the purpose to balance bed census. We constructively created a cyclical MSS with a length of four weeks. First, for each specialty, an integer number of OR blocks is chosen so that an output is achieved similar to the original MSS; due to this integrality average demand is slightly increased. Second, these blocks have been manually divided over the days in the MSS, and by trial-and-error a balanced outflow was realized.

As an illustration, Figure 7.2 displays the average bed utilization per weekday for care unit A (rejection probability <1%) before and after balancing the MSS. From this figure it is clear that both the midweek peak and the weekend dip can be cleared to a large extent, which results in distinct efficiency gains (see Table 7.3). We have reason to believe that even larger gains can be achieved. First, by developing a structured method to optimize the MSS instead of manual optimization. Second, the lack of detail in the available historical MSS data resulted in high variation in the input probability distributions of the number of cases per OR block and the length-of-stay distributions. When more information would be available on the content of MSS blocks, for instance on the level of subspecialty or even surgery type, the census predictions would show lower variability, resulting in lower bed requirements.

(5) Combination 1, 3, and 4. This intervention combines interventions (1), (3), and (4). Hospital management agreed upon a service level norm of rejection probabilities <2.5%. Under this requirement, it is possible to reduce the number of beds by 20% (from 104 to 83), and increase productivity by roughly 25%. Considering that the AMC has 30 inpatient departments, the savings potential for the entire hospital is substantial.

(6) Separation elective and acute. This intervention illustrates the capability of the model to provide quantitative support in decision making on care unit par-

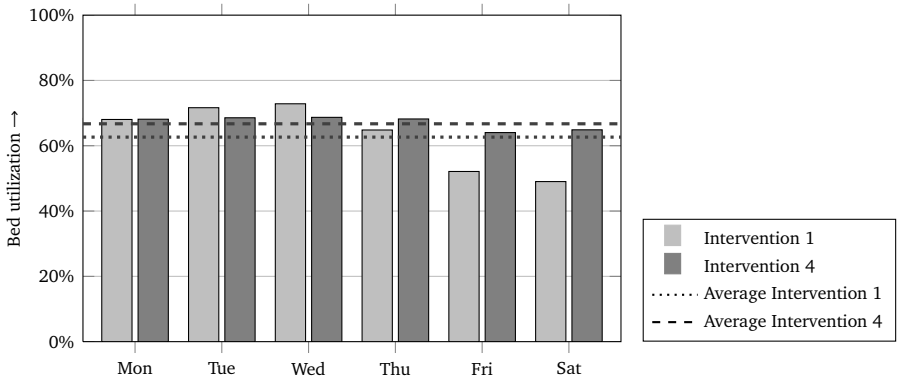


Figure 7.2: Average bed utilization per weekday (care unit A, rejections < 1%).

Table 7.4: The numerical results for intervention 6 (with the productivity-Δ% relative to 6a).

Intervention	*Unit*	*Capacity* (# beds)	*Rejection* (%)	*Misplace* (%)	*Utilization* (%)	*Floor*	*Capacity* (# beds)	*Productivity* (eq.(7.5))	(Δ%)
6a. Separation elective and acute									
Rejection < 5%	A	21	4.36	10.34	68.9	AB	42	45.0	-
	B	21	4.00	7.65	68.9				
	C	21	3.92	7.82	70.3	CD	42	57.6	-
	D	21	3.89	11.12	76.0				
Rejection < 2.5%	A	22	2.40	8.31	66.0	AB	44	43.7	-
	B	22	2.30	6.04	65.9				
	C	22	2.03	6.00	67.0	CD	44	56.1	-
	D	22	1.99	8.41	72.9				
Rejection < 1%	A	24	0.80	4.43	60.7	AB	47	41.6	-
	B	23	0.95	4.87	62.9				
	C	23	0.98	4.33	64.0	CD	46	54.2	-
	D	23	0.95	6.02	69.9				
6b. Combination (6a) and balance MSS									
Rejection < 5%	A	21	3.35	6.70	73.6	AB	41	48.9	+8.7
	B	20	3.30	8.55	75.7				
	C	21	3.92	7.82	70.3	CD	42	57.6	0.0
	D	21	3.89	11.12	76.0				
Rejection < 2.5%	A	22	2.07	4.33	70.6	AB	42	48.2	+10.3
	B	20	2.44	9.41	75.5				
	C	22	2.03	6.00	67.0	CD	44	56.1	0.0
	D	22	1.99	8.41	72.9				
Rejection < 1%	A	23	0.65	3.25	67.2	AB	45	45.8	+10.2
	B	22	0.59	3.90	69.2				
	C	23	0.98	4.33	64.0	CD	46	54.4	0.0
	D	23	0.95	6.02	69.9				
6c. Combination (6b) and change operational process									
Rejection < 5%	A	19	4.00	7.01	68.7	AB	39	51.3	+14.1
	B	20	3.07	4.46	66.4				
	C	20	4.73	9.59	71.8	CD	41	58.8	+2.0
	D	21	3.74	9.29	75.2				
Rejection < 2.5%	A	20	2.46	4.59	65.5	AB	40	50.6	+15.6
	B	20	2.38	5.16	66.2				
	C	22	2.10	4.62	65.9	CD	43	57.3	+2.2
	D	21	2.20	10.82	74.8				
Rejection < 1%	A	21	0.77	3.55	62.4	AB	43	47.9	+15.2
	B	22	0.56	2.04	60.3				
	C	23	0.78	3.60	62.6	CD	46	54.4	+0.2
	D	23	0.67	5.21	68.9				

titioning. Clinicians and managers in the AMC discuss the desirability to split elective and acute patient flows. Intervention 6a is formulated such that all elective patients are treated at Floor I (unit A: GEN, URO, VAS; unit B: PLA, TRA, ORT) and all acute patients at Floor II (unit C: GEN, URO, VAS, PLA; unit D:

TRA, ORT). In intervention 6b splitting electives and acute patients is combined with creating a balanced MSS, and intervention 6c extends this by including the changes in the operational process from intervention 3. Table 7.4 shows that the logistical performance is similar to the previous care unit configuration. We conclude therefore that whether or not to separate elective and acute patients in the studied case, should mainly be decided based on medical arguments.

7.5 Discussion

The design and operations of inpatient care facilities are typically to a large extent historically shaped. Accomplishing a better match with the changing environment is often possible, and even inevitable due to the pressure on hospital budgets. As an illustration, Dutch hospitals observe a shift from inpatient to outpatient care as a result of technological developments and increased medical knowledge. Consequently, many of these hospitals are organized in many care units that slowly decrease in size. Low bed utilizations occur, while at the same time a national shortage of nursing staff is observed. Therefore, the majority of Dutch hospitals is reorganizing its inpatient clinic. In this chapter, we have presented a generic analytical method that can support logistical decision-making for inpatient care services, by quantitatively predicting the impact of different scenarios and interventions.

We are able to assist decision-making on various planning levels. Insight can be gained on the impact of strategic (i.e., capacity dimensioning, case mix), tactical (i.e., the allocation of operating room time, misplacement rules), and operational decisions (i.e., time of admission/discharge). For these decisions, rules-of-thumb can be established. For example, we have shown the economies-of-scale effect: larger facilities can operate under a higher occupancy level than smaller ones in trying to achieve a given patient service level, since randomness balances out. In addition, by allowing overflow and setting appropriate rules, the benefits of bed capacity pooling are utilized, while the placement of patients on the preferred ward is maximized. Also, by adjusting the surgical schedule, extremely busy and quiet periods can be avoided. Once such basic rules are obtained, explicit interventions can be formulated of which the effect can be predicted. This combination between basic insights and quantifications is highly valuable to hospital managers.

The method is currently being applied in the AMC in redesigning its inpatient care services, of which the improvement potential is substantial (as numerically illustrated in this chapter). Such a process of drastically changing an existing healthcare environment is highly political. We believe that the benefit of quantitative analysis in such a 'negotiation' process is that it rationalizes the process of realizing a good trade-off between interests of clinicians and patients. Quantification ensures that robust organizational plans are formulated, for instance also anticipating for the expected increase of acute admissions due to a changing nature of the emergency department. Finally, we observe that applying the method and discussing the results triggers the discussion to also focus on other potential gains like a more efficient use of the operating theater.

In follow-up research we focus on three directions. First, by combining the current inpatient census with the settled upcoming MSS, the model could be exploited to support last minute decision-making like whether or not to hire temporary staff. Second, we will focus on incorporating the possibility of intermediate intensive care unit stays for patients who have undergone a complex surgery. Finally, the hourly level of the model will provide the basis for a formal approach along which effective and efficient nurse staffing can be achieved. The latter will be the focus of the next chapter.

7.6 Appendix

In the appendix, the derivations are presented that were omitted in the main text for reasons of readability. The exposition is such that it is supplementary to the main text, and is therefore not intended to be comprehensible in isolation.

7.6.1 Demand predictions for elective patients

Single surgery block. To calculate $a^j_{n,t}(x)$, we first determine the admission process under a given number of performed surgeries y. Define $a^j_{n,t}(x|y)$ as the probability that x patients are admitted until time t on day n, given that y admissions take place in total. Then, $a^j_{n,t}(x)$ is calculated by

$$a^j_{n,t}(x) = \sum_{y=x}^{C^j} a^j_{n,t}(x|y)c^j(y), \tag{7.6}$$

With $v^j_{n,t}$ the probability for a type j patient to be admitted in time t, given that he/she will be admitted at day n and is not yet admitted before t:

$$v^j_{n,t} = \frac{w^j_{n,t}e^j_n}{e^j_n \sum_{k=t}^{T-1} w^j_{n,k} + e^j_0 \cdot \mathbb{1}_{(n=-1)}},$$

in (7.6), $a^j_{n,t}(x|y)$ is calculated as follows. For $n=-1, t=0$,

$$a^j_{n,t}(x|y) = \binom{y}{x}(v^j_{n,t})^x(1-v^j_{n,t})^{y-x},$$

for $n=0, t=0$,

$$a^j_{n,t}(x|y) = \sum_{g=0}^{x}\binom{y-g}{x-g}(v^j_{n,t})^{x-g}(1-v^j_{n,t})^{y-x}a^j_{n-1,T-1}(g|y),$$

for $n=-1, t=1,\dots,T-1$, and for $n=0, t=1,\dots,\vartheta_j-1$,

$$a^j_{n,t}(x|y) = \sum_{g=0}^{x}\binom{y-g}{x-g}(v^j_{n,t})^{x-g}(1-v^j_{n,t})^{y-x}a^j_{n,t-1}(g|y),$$

and for $n = 0, t \geq \vartheta_j$,

$$a^j_{n,t}(x|y) = 0.$$

To calculate $d^j_{n,t}(x)$, we first determine $d^j_n(x)$, for day 0 the probability that x patients are present at the start of the discharge process ($t = \vartheta_j$) and for days $n > 0$ the probability that x patients are present at the start of the day:

$$d^j_n(x) = \begin{cases} c^j(x) & \text{, if } n = 0, \\ \sum_{g=x}^{C^j} \binom{g}{x} (s^j_{n-1})^{g-x} (1 - s^j_{n-1})^x d^j_{n-1}(g) & \text{, if } n = 1, \ldots, L^j, \end{cases}$$

where s^j_n is the probability that a type j patient who is still present at the begin of day n is discharged on day n:

$$s^j_n = \frac{P^j(n)}{\prod_{m=0}^{n-1}(1 - s^j_m)}.$$

Starting from $d^j_n(x)$, we determine the day process. For $n = 0, t < \vartheta_j$,

$$d^j_{n,t}(x) = 0,$$

for $n = 0, t = \vartheta_j$ and for $n > 0, t = 0$,

$$d^j_{n,t}(x) = d^j_n(x),$$

and for $n = 0, t > \vartheta_j$, and for $n > 0, t > 0$,

$$d^j_{n,t}(x) = \sum_{k=x}^{C^j} \binom{k}{x} (z^j_{n,t-1})^{k-x} (1 - z^j_{n,t-1})^x d^j_{n,t-1}(k),$$

where $z^j_{n,t}$ is the probability of a type j patient to be discharged during time interval $[t, t+1)$ on day n, given this patient is still present at time t:

$$z^j_{n,t} = \frac{m^j_{n,t} P^j(n)}{P^j(n) \sum_{i=t}^{T-1} m^j_{n,i} + \sum_{k=n+1}^{L^j} P^j(k)}.$$

Single MSS cycle. We determine the overall probability distribution of the number of patients in recovery resulting from a single MSS, using discrete convolutions. If specialty j is assigned to OR block $b_{i,s}$, then the distribution $\bar{h}^{i,s}_{m,t}$ for the number of recovering patients of block $b_{i,s}$ present at time t on day m ($m \in \{0, 1, 2, \ldots, S, S+1, S+2, \ldots\}$) is given by:

$$\bar{h}^{i,s}_{m,t} = \begin{cases} \mathbf{0} & \text{, if } m < s - 1, \\ h^j_{m-s,t} & \text{, if } m \geq s - 1, \end{cases}$$

where $\mathbf{0}$ means $\bar{h}^{i,sj}_{m,t}(0) = 1$ and all other probabilities $\bar{h}^{i,s}_{m,t}(x), x > 0$ are 0. Then, $H_{m,t}$ is computed by:

$$H_{m,t} = \bar{h}^{1,1}_{m,t} \otimes \bar{h}^{1,2}_{m,t} \otimes \ldots \otimes \bar{h}^{1,S}_{m,t} \otimes \bar{h}^{2,1}_{m,t} \otimes \ldots \otimes \bar{h}^{I,S}_{m,t}. \tag{7.7}$$

Steady state. Since the cyclic structure of the MSS implies that the recovery of patients receiving surgery during one cycle may overlap with patients from the next cycle, the distributions $H_{m,t}$ have to be overlapped in the correct manner. $H^{SS}_{s,t}$ can be computed as follows:

$$H^{SS}_{s,t} = \begin{cases} H_{s,t} \otimes H_{s+S,t} \otimes \ldots \otimes H_{s+\lceil M/S \rceil S,t} & \text{, if } s = 1, \ldots, S-1, \\ H_{0,t} \otimes H_{S,t} \otimes \ldots \otimes H_{\lceil M/S \rceil S,t} & \text{, if } s = S. \end{cases}$$

where $M = \max\{m \mid \exists t, x \text{ with } H_{m,t}(x) > 0\}$.

7.6.2 Demand predictions for acute patient types

Single patient type. For patient type $j = (p, r, \theta)$, the admission process $\tilde{a}^j_t$ is determined by a non-homogeneous Poisson process:

$$\tilde{a}^j_t(x) = \frac{(\lambda^j)^x e^{-\lambda^j}}{x!}, \quad t = \theta.$$

To calculate $\tilde{d}^j_{n,t}(x)$, we first determine $\tilde{d}^j_n(x)$, for day 0 the probability that x patients are present at the start of the discharge process ($t = \theta + 1$) and for days $n > 0$ the probability that x patients are present at the start of the day:

$$\tilde{d}^j_n(x) = \begin{cases} \tilde{a}^j_\theta(x) & \text{, if } n = 0, \\ \sum_{g=x}^{\infty} \binom{g}{x} (\tilde{s}^j_{n-1})^{g-x} (1 - \tilde{s}^j_{n-1})^x \tilde{d}^j_{n-1}(g) & \text{, if } n = 1, \ldots, L^j, \end{cases}$$

where $\tilde{s}^j_n$ is the probability that a type j patient who is still present at the begin of day n is discharged during day n:

$$\tilde{s}^j_n = \frac{P^j(n)}{\prod_{m=0}^{n-1}(1 - \tilde{s}^j_m)}.$$

Starting from $\tilde{d}^j_n$, we determine the day process. For $n = 0, t \leq \theta$,

$$\tilde{d}^j_{n,t}(x) = 0,$$

for $n = 0, t = \theta + 1$, and for $n > 0, t = 0$,

$$\tilde{d}^j_{n,t}(x) = \tilde{d}^j_n(x),$$

and for $n = 0, t > \theta + 1$, and for $n > 0, t > 0$,

$$\tilde{d}^j_{n,t}(x) = \sum_{k=x}^{\infty} \binom{k}{x} (\tilde{z}^j_{n,t-1})^{k-x} (1 - \tilde{z}^j_{n,t-1})^x \tilde{d}^j_{n,t-1},$$

where $\tilde{z}^j_{n,t}$ is the probability of a type j patient to be discharged during time interval $[t, t+1)$ on day n, given this patient is still present at time t:

$$\tilde{z}^j_{n,t} = \frac{\tilde{m}^j_{n,t} P^j(n)}{P^j(n) \sum_{i=t}^{T-1} \tilde{m}^j_{n,i} + \sum_{k=n+1}^{L^j} P^j(k)}.$$

Single cycle. To determine the overall probability distribution of the number of patients in recovery resulting from a single AAC, define $\bar{g}^j_{w,t}$ as the probability distribution of the number of recovering patients of type j present at time interval t on day w ($w \in \{0, 1, 2, \ldots, R, R+1, R+2, \ldots\}$). The distribution $\bar{g}^j_{w,t}$ is given by:

$$\bar{g}^j_{w,t} = \bar{g}^{p,r,\theta}_{w,t} = \begin{cases} \mathbf{0} & \text{, if } w < r, \\ g^j_{w-r,t} & \text{, if } w \geq r. \end{cases}$$

Then, $G_{w,t}$ is computed by:

$$G_{w,t} = \bar{g}^{1,1,0}_{w,t} \otimes \ldots \otimes \bar{g}^{1,1,T-1}_{w,t} \otimes \bar{g}^{1,2,0}_{w,t} \otimes \ldots \otimes \bar{g}^{1,2,T-1}_{w,t} \otimes \bar{g}^{2,1,0}_{w,t} \otimes \ldots \otimes \bar{g}^{P,R,T-1}_{w,t}. \quad (7.8)$$

Steady state. $G^{SS}_{r,t}$ can be computed as follows:

$$G^{SS}_{r,t} = G_{r,t} \otimes G_{r+R,t} \otimes G_{r+2R,t} \otimes \ldots \otimes G_{r+\lceil W/R \rceil R,t},$$

where $W = \max\{r \mid \exists t, x \text{ with } G_{r,t}(x) > 0\}$.

7.6.3 Performance indicators

In this appendix, the derivation of $Z^k_{q,t}(x_k|n)$ is presented. To this end, let us first introduce the concept *cohort*. A cohort is a group of patients originating from a single instance of an OR block (electives) or admission time interval (acute patients). Then,

$$\begin{aligned} Z^k_{q,t}(x_k|n) &= \frac{P\left[\begin{array}{c}\text{Demand } x_k \text{ patients for unit } k \text{ on time } t \text{ on} \\ \text{day } q \text{ of which } n \text{ are arriving in } [t, t+1)\end{array}\right]}{P[n \text{ arrivals for unit } k \text{ on day } q \text{ in } [t, t+1)]} \\ &= \frac{1}{\Lambda^k_{q,t}(n)} \sum_{\substack{y_{\sigma(1)},\ldots,y_{\sigma(\Omega)}, \\ n_{\sigma(1)},\ldots,n_{\sigma(\omega)}: \\ \sum_i y_i = x_k, \sum_j n_j = n}} \left\{ \prod_{i=\omega+1}^{\Omega} f^{\sigma(i)}_{q,t}(y_{\sigma(i)}) \right\} \cdot \\ &\qquad \left\{ \prod_{j=1}^{\omega} \alpha^{\sigma(j)}_{q,t}(y_{\sigma(j)}) \breve{a}^{\sigma(j)}_{q,t}(n_{\sigma(j)}|y_{\sigma(j)}) \right\}, \end{aligned}$$

where Ω is the total number of cohorts, ω the number of cohorts that do generate arrivals during time interval $[t, t+1)$ on day q, and the permutation σ is such that the patient types $\sigma(1), \ldots, \sigma(\omega)$ are the types that can generate those arrivals. Further, for notational convenience we introduce the function $f^i_{q,t}$ as $f^i_{q,t} = h^i_{q,t}$ for the elective patients, and $f^i_{q,t} = g^i_{q,t}$ for acute patient types. Also, we introduce $\alpha^j_{q,t}$ as $\alpha^j_{q,t} = a^j_{q,t}$ for the elective patient types, and

$$\alpha^j_{q,t} = \tilde{a}_t^{(p,q \bmod R + R \cdot \mathbb{1}_{q \bmod R = 0}, t)},$$

for the acute patient types. It remains to define $\check{a}^j_{q,t}(n_j|y_j)$, the probability that for an arriving cohort, from the y_j patients present in total, n_j arrivals occur during time interval $[t, t+1)$:

$$\check{a}^j_{q,t}(n_j|y_j) = \binom{y_j}{n_j} (\nu^j_{n,t})^{n_j} (1 - \nu^j_{n,t})^{y_j - n_j},$$

where for elective patient types

$$\nu^j_{n,t} = \frac{w^j_{n,t} e^j_n}{e^j_n \sum_{k=0}^{t} w^j_{n,k} + e^j_{-1} \cdot \mathbb{1}_{(n=0)}}$$

and for acute patient types $\nu^j_{n,t} = 1$.

CHAPTER 8

Flexible Nurse Staffing

8.1 Introduction

Deploying adequate nurse staffing levels is one of the prime responsibilities of inpatient care facility managers. Nursing staff typically accounts for the majority of hospital budgets [644], which makes that every appearance of overstaffing is scrutinized in times that tight cost-containment efforts are required [377]. At the same time, maintaining appropriate staffing levels is crucial to be able to provide high-quality care. There is a growing body of evidence implicating associations between decreased staffing and higher hospital related mortality and adverse patient events [341, 453], and increased work stress and burnout among nurses [9, 10]. In this chapter, we present an exact method to assist healthcare administrators in ensuring safe patient care, while also maintaining an efficient and cost-effective nursing service.

Workload on nursing wards depends highly on patient arrivals and lengths of stay, which are both inherently variable. Predicting workload, and staffing nurses accordingly, is essential for guaranteeing quality of care in a cost effective manner [85, 602]. Accurate workload predictions require the consideration of the dynamics of surrounding departments, since many patient arrivals at the inpatient care facility originate from the operating theater and the emergency department. In Chapter 7, we presented a method to predict bed census by hour in various care units of an inpatient clinic as a function of the operating room block schedule and a cyclic arrival pattern of emergency patients. The stochastic analytic model presented in the current chapter takes these predictions of Chapter 7 as starting point to determine appropriate nurse staffing levels.

Chapter 7 emphasized the importance of recognizing the interrelation between various planning decisions when designing and operating inpatient care services. Presented examples of decisions for which it is desirable to be made in coordination with each other were case mix, care unit partitioning, and care unit size. In addition, especially for surgical inpatient departments, alignment with the planning of the operating room schedule was shown to be beneficial. All these decisions are also intertwined with inpatient care workforce requirements, for example in terms of skill mix, number of full time equivalents, and staffing levels per working shift. In the current chapter, we focus on incorporating the tactical decision in Chapter 2

referred to as 'staff-shift scheduling' in our integrated modeling framework. We address the question: for all working shifts during the planning horizon, how many employees should be assigned to each inpatient care unit? These numbers, in turn, provide a guideline for the workforce dimensioning decision on the strategic level.

We explore the potential of flexible staffing policies that allow hospitals to dynamically respond to their fluctuating patient population. This flexibility is achieved by employing a pool of cross-trained nurses, for whom it is only at the start of a shift decided in which specific care units they will work. The commonly applied term for such flexible employees is 'float nurses' [236, 540]. The basic idea behind the possible added value of the introduction of flex pools is the following. Although the inpatient population fluctuates, this fluctuation is to a certain extent predictable, due to its dependance on the operating room schedule and other predictable variability in patient arrivals (e.g., seasonality, day-of-week, and time-of-day effects). This predictable variation can be taken into account when setting the staffing levels for 'dedicated nurses', nurses with a fixed assignment to a care unit. Typically, staffing levels are to be determined a number of weeks in advance, so that individual nurse rosters can be settled timely. Therefore, when only dedicated nurses are employed, the buffer capacity required to protect against random demand fluctuations can lead to regular overstaffing. When two or more care units cooperate by jointly appointing a flexible nurse pool, the variability of these random demand fluctuations balances out due to economies of scale, so that less buffer capacity is required.

Nurse-to-patient ratios are commonly used when determining staffing levels [10, 650]. These ratios indicate how many patients a registered nurse can care for during a shift, taking into account both direct and indirect patient care. Staffing according to nurse-to-patient ratios can be done in two ways. The ratios can be taken as mandatory lower bound, such as in California (USA) and Victoria (Australia), where legal minimums for nurse-to-patient ratios were set for general medical and surgical wards [11, 573]. The advantage of such minimum ratios is that a consistently high level of patient safety is guaranteed [341, 377]. The disadvantage, however, is that always all beds need to be staffed, because the possibility that all beds are occupied always exists and, as described, the nurse rosters have to be settled in advance. Therefore, overstaffing is a threat, since there is no flexibility to adjust staffing levels to predicted patient demand. Overcoming this disadvantage, a second version of applying nurse-to-patient ratios exists, which involves taking them merely as guidelines [192]. In that case, the assumption is that there exists slack in the time window within which some indirect patient care tasks can be performed, without having direct negative consequences on patient safety or work stress. As a result, the ratios may sometimes be violated, but not too often, and not too long. In our approach, we combine the advantages of both approaches, by utilizing two nurse-to-patient ratio targets. The first ratio needs to be satisfied at all times, while the second more restrictive ratio only for a certain fraction of time.

Our contribution is a generic exact analytic approach to find the number of nurses to be staffed each working shift that guarantees a desired quality of care reflected by nurse-to-patient ratios, in the most cost-effective manner. The approach

directly connects with the bed census prediction method presented in Chapter 7, so that alignment of staffing decisions with other interrelated inpatient planning decisions can be achieved, as well as coordination with the operating theater and the emergency department. First, to match nursing capacity with demand predictions, a stochastic mathematical program is formulated to determine optimal staffing levels when only dedicated nurses are employed: the 'fixed staffing policy' model. Next, we formulate a model in which the flex pool with float nurses is introduced, and in which exactly the same quality constraints are satisfied as in the fixed staffing policy model. The formulation of the flexible staffing policy model includes an assignment procedure that prescribes the rules according to which the float nurses are assigned to specific care units at the start of each working shift. Because the flexible staffing model is computationally too expensive to solve to optimality in reasonable time, we present an approximation model, which provides a lower and an upper bound on the staffing requirements.

To illustrate its potential, the method is applied to the same case study as that of Chapter 7. This case study involves the care units in the surgical inpatient clinic of the Academic Medical Center (AMC) Amsterdam, serving the specialties traumatology, orthopedics, plastic surgery, urology, vascular surgery, and general surgery. Inspired by the numerical results, the AMC decided that the flexible nurse staffing method will be fully implemented during the upcoming years, as part of the total redesign of its inpatient care services.

The chapter is organized as follows. Section 8.2 provides a review of relevant literature. Section 8.3 presents the staffing models for the fixed and the flexible staffing policies. Section 8.4 presents the numerical results, and Section 8.5 closes the chapter with a general discussion.

8.2 Background: workforce planning

Capacity planning for nursing staff has received considerable attention from the operations research community. The nurse staffing process involves a set of hierarchical decisions over different time horizons with different precision. The interdependence of the decision levels must be recognized to bring about systematic nurse staffing improvements. As expressed in the literature review [483], each level is constrained by previous commitments made at higher levels, and by the degrees of flexibility for later correction at lower levels. For a more elaborate exposition of the relevant decisions and considerations involved at each decision level, and a detailed overview of relevant literature, we refer the reader to Chapter 2.

The literature has mainly focused on nurse rostering, for example reflected by the survey and classification articles [91, 150, 197]. Although the rostering methods are computationally efficient and very helpful to support practitioners in creating timetables, they generally take required staffing levels as prerequisite information [77, 280]. Incorrect assumptions on the (tactical) required staffing levels, during the (operational offline) rostering process, might therefore result in the necessity to make expensive corrections on the operational online decision level, for instance

by additionally hiring temporary staff. Therefore, to be able to provide adequate input for the rostering process, we focus on the tactical decision level, by specifying appropriate 24-hours-a-day-staffing levels divided in shifts (e.g., a day, evening and night shift).

Tactical workforce decision making in healthcare has received less attention. A spreadsheet approach has been presented in [192], to retrospectively fit optimal shift staffing levels on historical census data. Prospectively assessing the impact of alternative interventions is difficult via such approaches, since they lack the flexibility to explicitly model and study the coordination between different inpatient care decision levels, and the alignment with surrounding departments. Simulation studies have shown to be successful in taking a more integral approach (e.g., [261, 280]). The inherent disadvantage of simulation studies is, however, that they are typically context-specific, which limits the generalizability of study outcomes. Analytic, but deterministic, approaches can for example be found in [41, 461, 618]. Stochastic approaches to determine shift staffing levels are available in [602, 644, 650]. None of these references take an integral approach, as the demand distributions underlying the staffing decisions are not based on patient arrival patterns from the operating theater and emergency department.

Workforce flexibility is indicated as a powerful concept in reducing the required size of workforce and increasing job satisfaction [91, 152, 236, 261, 329, 534, 540]. To adequately respond to patient demand variability various types of flexibility are suggested, among which the use of part-time employees, overtime, temporary agency employees, and float nurses. Related to our work are the articles [236, 392] in which the potential of float pools with cross-trained nurses is investigated. Both these references address the aggregate decision which budget of float nurse hours should be available during a given time horizon, and, as such, do not address the level of working shifts. For the assignment strategy of a given number of available float nurses to care units at the start of working shifts the authors of [569] indicate that formulating such an assignment strategy requires the consideration of three issues: (1) a methodology for the measurement of the severity of need for an additional nurse, (2) a prediction per care unit of that severity of need during an upcoming shift, and (3) development of a technique for the allocation of the available float nurses to care units to meet this need. While [569] focuses on the third issue by developing a branch-and-bound algorithm, our assignment strategy involves the consideration of all three steps.

Staffing according to nurse-to-patient ratios has received attention in the operations research literature in [602, 644, 650]. Both [602] and [644] indicate that in practice, setting the numerical values of the ratios is more based on negotiation than on science. The authors of [644] studied the relation between staffing costs and nurse-to-patient ratios. In this article, also two interesting directions for future research were stated: first, exploring the use of float nurse pools in satisfying nurse-to-patient ratios, and, second, developing models to make scientific recommendations on the numerical values of the ratios. The first issue is addressed in the current study. The second issue has been the focus of [602, 650]. Both these references

present a queueing model along which they motivate that the ratios as mandated in California are too rigid. They underline the importance of differentiating ratios with patient mix (reflecting the severity of patients' illnesses and their acuity), and with care unit size. In our study, we focus on determining staffing levels given pre-specified nurse-to-patient ratios. Nevertheless, we do want to stress the importance of employing meaningful nurse-to-patient ratios in realizing high-quality staffing.

To conclude, our contribution is an exact stochastic analytic approach, aimed at deriving appropriate staffing levels, including the flexibility of float nurses, using nurse-to-patient ratios, while taking an integrated care chain perspective.

8.3 Methods

In this section, the staffing models are presented. The staffing models are based on the bed census predictions that are obtained from the model of Chapter 7. Section 8.3.1 discusses the requirements that need to be satisfied in setting appropriate staffing levels. Section 8.3.2 presents the fixed staffing model. Section 8.3.3 formulates the model to find optimal staffing levels when float nurse pools are applied: the flexible staffing model. Since the flexible model suffers from the curse of dimensionality, we approximate the solution via two models that respectively find upper and lower bounds on the staffing requirements.

8.3.1 Staffing requirements

We consider a planning horizon of Q days ($q = 1, \ldots, Q$). Each day is divided in T time intervals ($t = 0, 1, \ldots, T-1$). The set of working shifts is denoted by $\mathcal{T}$, where a shift τ is characterized by its start time b_τ and its length ℓ_τ. Within the time horizon (q, t) is a unique time interval and (q, τ) a unique shift. For notational convenience, $t \geq T$ indicates a time interval on a later day, e.g., $(q, T+5) = (q+1, 5)$. For each of K inpatient care units, with the capacity of unit k being M^k beds, staffing levels have to be determined for each shift (q, τ).

We consider two types of staffing policies: 'fixed' and 'flexible' staffing. Under fixed staffing the number of nurses working in unit k during shift (q, τ), denoted by $s^k_{q,\tau}$, is completely determined in advance. In the flexible case, 'dedicated' staffing levels $d^k_{q,\tau}$ per unit are determined, together with a number of nurses $f_{q,\tau}$ available in a flex pool. The decision to which particular units the float nurses are assigned is delayed until the start of the execution of a shift. We assign float nurses to one and the same care unit for a complete working shift, to avoid many hand-overs, which increase the risk of medical errors. Thus, we obtain staffing levels $s^k_{q,\tau} = d^k_{q,\tau} + f^k_{q,\tau}$, $k = 1, \ldots, K$, where $f^k_{q,\tau}$ denotes the number of float nurses assigned to unit k from the available $f_{q,\tau}$. Taking into account the current bed census and the predictions on patient admissions and discharges, the allocation of the float nurses to care units at the start of a shift is done according to a predetermined assignment procedure. We denote such an assignment procedure by π.

Our goal is to determine the most cost-efficient staffing levels such that certain quality-of-care constraints are satisfied. Since float nurses are required to be cross-trained it is likely that these are more expensive. To be able to differentiate, we therefore consider staffing costs ω_d for each dedicated nurse that is staffed for one shift and ω_f for each flexible nurse. Next, the nurse-to-patient ratio targets during shift (q,τ) are reflected by $r^k_{q,\tau}$, indicating the number of patients a nurse can be responsible for at any point in time. To keep track of the compliance to these targets, we define the concept 'nurse-to-patient coverage', or shortly 'coverage'. With x^k the number of patients present at unit k at a certain time (q,t), $b_\tau \leq t < b_\tau + \ell_\tau$, the coverage is given by $r^k_{q,\tau} \cdot s^k_{q,\tau}/x^k$. Thus, a coverage of one or higher corresponds to the preferred situation.

Starting from the following quality-of-care requirements as prerequisites, we will formulate the fixed and flexible staffing models by which the most cost-effective staffing levels can be found:

(i) **Staffing minimum.** For safety reasons, at least S^k nurses have to be present at care unit k at any time.

(ii) **Coverage minimum.** The coverage at care unit k may never drop below β^k.

(iii) **Coverage compliance.** The long-run fraction of time that the coverage at care unit k is one or higher is at least α^k. We denote the expected compliance at care unit k during shift (q,τ) by $c^k_{q,\tau}(\cdot)$; the arguments of this function depend on which staffing policy is considered.

(iv) **Flexibility ratio.** To ensure continuity of care, at any time, the fraction of nurses at care unit k that are dedicated nurses has to be at least γ^k.

(v) **Fair float nurse assignment.** The policy π, according to which the allocation of the available float nurses to care units at the start of a shift is done, has to be 'fair'. Fair is defined as assigning every next float nurse to the care unit where the expected coverage compliance during the upcoming shift is the lowest.

8.3.2 Fixed staffing

When only dedicated staffing is allowed, there is no interaction between care units. Therefore, the staffing problem decomposes in the following separate decision problems for each care unit k, and each shift (q,τ):

$$\min \quad z_F = \quad \omega_d s^k_{q,\tau} \tag{8.1}$$

$$\text{s.t.} \quad s^k_{q,\tau} \geq S^k \tag{8.2}$$

$$s^k_{q,\tau} \geq \left\lceil \beta^k \cdot M^k / r^k_{q,\tau} \right\rceil \tag{8.3}$$

$$c^k_{q,\tau}\left(s^k_{q,\tau}, r^k_{q,\tau}\right) \geq \alpha^k \tag{8.4}$$

The constraints (8.2), (8.3), and (8.4) reflect requirements (i), (ii), and (iii), respectively. Let $X^k_{q,t}$ be the random variable with bed census distribution $\hat{Z}^k_{q,i}$ counting the

number of patients present on care unit k at time (q,t). Then, the coverage compliance in (8.4) can be calculated as follows:

$$\begin{aligned} c^k_{q,\tau}\big(s^k_{q,\tau}, r^k_{q,\tau}\big) &= \mathbb{E}\left[\frac{1}{\ell_\tau}\sum_{t=b_\tau}^{b_\tau+\ell_\tau-1}\mathbb{1}\big(X^k_{q,t} \le s^k_{q,\tau}\cdot r^k_{q,\tau}\big)\right] \\ &= \frac{1}{\ell_\tau}\sum_{t=b_\tau}^{b_\tau+\ell_\tau-1}\sum_{x=0}^{s^k_{q,\tau}\cdot r^k_{q,\tau}}\hat{Z}^k_{q,t}(x). \end{aligned}$$

Observe that the term $\sum_{x=0}^{s^k_{q,\tau}\cdot r^k_{q,\tau}}\hat{Z}^k_{q,t}(x)$ reflects the probability that with staffing level $s^k_{q,\tau}$ and under ratio $r^k_{q,\tau}$ the nurse-to-patient ratio target is satisfied during time interval $[t,t+1)$. The optimum of (8.1) is found by choosing the minimum $s^k_{q,\tau}$ satisfying constraints (8.2) and (8.3), and increasing it until constraint (8.4) is satisfied.

8.3.3 Flexible staffing

The next step is to formulate the flexible staffing model. Note that for requirements (i) and (ii), the constraints are similar to those for fixed staffing. Under the assumption $\omega_d \le \omega_f$, we can replace $s^k_{q,\tau}$ by $d^k_{q,\tau}$ in (8.2) and (8.3). Due to the presence of a flex pool the care units cannot be considered in isolation anymore. Hence, constraint (8.4) has to be replaced. An assignment procedure has to be formulated that fulfils requirement (v), and this assignment procedure influences the formulation of the constraint for requirement (iii). In addition, a constraint needs to be added for requirement (iv).

For an assignment procedure π that allocates the float nurses to care units at the start of a shift (q,τ), let $g^\pi_{q,\tau}(\boldsymbol{d}, f, \boldsymbol{y})$ be the vector of length K denoting the number of float nurses assigned to each care unit, when f flex nurses are available to allocate, the number of staffed dedicated nurses equals $\boldsymbol{d} = (d^1,\ldots,d^K)$, and the census at the different care units at time (q,b_τ) equals $\boldsymbol{y} = (y^1,\ldots,y^K)$. A vector of the type $\boldsymbol{y}$ reflects what we will call a *census configuration*.

Let π^* denote the assignment procedure that ensures constraint (v). The assignment procedure π^* depends on $\boldsymbol{d}_{q,\tau}$, $f_{q,\tau}$, and $r^k_{q,\tau}, k = 1,\ldots,K$, and therefore also the coverage does. Hence, requirement (v) gives a constraint of the form $c^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}) \ge \alpha^k$. But, in addition, assignment procedure π^* depends on the census configuration $\boldsymbol{y}$ at time (q,b_τ), so to be able to calculate the coverage compliance we first need to compute $c^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}; \boldsymbol{y})$, the coverage compliance given that at the start of shift (q,τ) census configuration $\boldsymbol{y}$ is observed. Then, the coverage compliance is given by:

$$c^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}) = \sum_{\boldsymbol{y}}\Big\{c^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}; \boldsymbol{y})\prod_{w=1}^{K}\hat{Z}^w_{q,\tau}(y^w)\Big\}.$$

Using $c^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}; \boldsymbol{y})$, the assignment policy π^* satisfying requirement (v) is

the one that satisfies:

$$g^{\pi^*}_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, \boldsymbol{y}) = \max_{\left\{f^1_{q,\tau},\ldots,f^K_{q,\tau} : \sum_k f^k_{q,\tau} = f_{q,\tau}\right\}} \min_k \; c^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}; \boldsymbol{y}). \tag{8.5}$$

Applying policy π^* provides $s^k_{q,\tau}(\boldsymbol{y})$, the number of nurses staffed at care unit k if census configuration $\boldsymbol{y}$ is observed at the start of shift (q,τ). Hence, the flexible model is, for each shift (q,τ):

$$\min \quad z_E = \omega_f f_{q,\tau} + \sum\nolimits_k \omega_d d^k_{q,\tau} \tag{8.6}$$

$$\text{s.t.} \quad d^k_{q,\tau} \geq S^k \qquad \text{, for all } k, \tag{8.7}$$

$$d^k_{q,\tau} \geq \left\lceil \beta^k \cdot M^k / r^k_{q,\tau} \right\rceil \qquad \text{, for all } k, \tag{8.8}$$

$$c^k_{q,\tau}(\boldsymbol{d}_{q,t}, f_{q,\tau}, r^k_{q,\tau}) \geq \alpha^k \qquad \text{, for all } k, \tag{8.9}$$

$$d^k_{q,\tau} \geq \gamma^k \cdot s^k_{q,\tau}(\boldsymbol{y}) \qquad \text{, for all } k, \boldsymbol{y}, \tag{8.10}$$

$$s^k_{q,\tau}(\boldsymbol{y}) = d^k_{q,\tau} + g^{k,\pi^*}_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, \boldsymbol{y}) \qquad \text{, for all } k, \boldsymbol{y}. \tag{8.11}$$

Constraints (8.7)–(8.11) reflect (i)–(v) respectively. Finding the optimum for (8.6) requires the computation of $c^k_{q,\tau}(\boldsymbol{d}, f_{q,\tau}, r^k_{q,\tau}; \boldsymbol{y})$ by considering every sample path of census configurations during a shift. For realistic instances this is computationally too expensive to find the optimal solution for $d^1_{q,\tau}, \ldots, d^K_{q,\tau}, f_{q,\tau}$ in a reasonable amount of time (see Appendix 8.6.1). Therefore, two approximations are proposed. The first approximation is obtained by deriving the probability distribution for the maximum number of patients present during each shift, and finding the optimal staffing for this maximum census. In this case the number of patients present is overestimated, therefore the required staffing levels are overestimated, and thus we obtain an upper bound on the staffing requirements. In the second approximation we reassign the float nurses to the care units at the start of each time interval. Since this provides more flexibility to align the float nurse allocation to the current census, we obtain an underestimation of the required staffing levels. As such, a lower bound on the actual staffing requirements is found. Finally, comparing the lower and upper bound solutions and the solution for the fixed model, provides us (an approximation of) the optimal solution of the flexible staffing model. To be more specific, the upper bound solution guarantees that the constraints are satisfied in the flexible staffing model. When the lower bound solution coincides with the upper bound or the fixed staffing solution, we are sure to have found the optimal solution. Otherwise the lower bound provides an error bound.

Upper bound model. Based on the observed maximum census configuration $\boldsymbol{x} = (x^1, \ldots, x^K)$ during a shift, let π^{up} be the assignment policy that allocates the nurses from the flex pool to the care units where the number of nurses short is the highest:

$$g^{\pi^{up}}_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, \boldsymbol{x}) = \max_{\left\{f^1_{q,\tau},\ldots,f^K_{q,\tau} : \sum_k f^k_{q,\tau} = f_{q,\tau}\right\}} \min_k \; \frac{r^k_{q,\tau} \cdot (d^k_{q,\tau} + f^k_{q,\tau}) - x^k}{r^k_{q,\tau}}.$$

Let $\hat{W}^k_{q,\tau}(x)$ be the probability that during shift (q,τ) the maximum census level that occurs at care unit k is x patients. These probabilities are derived by analogy with the derivation of $\hat{Z}^k_{q,\tau}(x)$ in Chapter 7 (for details see Appendix 8.6.2). To obtain the upper bound, for $b_\tau \leq t < b_\tau + \ell_\tau$, we approximate the original distribution $\hat{Z}^k_{q,t}(x)$ by $\hat{W}^k_{q,\tau}(x)$. Let $\bar{X}^k_{q,\tau}$ be the random variable with distribution $\hat{W}^k_{q,\tau}$ that counts the maximum number of patients on care unit k during working shift (q,τ). To see that this approximation leads to an upper bound on the required staffing levels, observe that $\bar{X}^k_{q,\tau} \geq X^k_{q,t}$, for $b_\tau \leq t < b_\tau + \ell_\tau$, so that for every time interval of a shift the census is overestimated, and thus staffing requirements are overestimated.

Since we use the same census distribution in every time interval during a shift, the coverage compliance over a shift $\bar{c}^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau})$ is calculated by:

$$\bar{c}^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}) = \sum_{\boldsymbol{x}} \Big\{ \mathbb{1}\big(x^k \leq r^k_{q,\tau} \cdot s^k_{q,\tau}(\boldsymbol{x})\big) \cdot \prod_{w=1}^{K} \hat{W}^w_{q,\tau}(x^w) \Big\},$$

where $s^k_{q,\tau}(\boldsymbol{x})$ is the number of nurses staffed at care unit k for shift (q,τ) under assignment policy π^{up}, when the maximum observed census configuration is $\boldsymbol{x}$. Summarizing, for each shift (q,τ), we have:

$$\min \quad z_U = \omega_f f_{q,\tau} + \sum\nolimits_k \omega_d d^k_{q,\tau} \tag{8.12}$$

$$\text{s.t.} \quad d^k_{q,\tau} \geq S^k \qquad \text{, for all } k, \tag{8.13}$$

$$d^k_{q,\tau} \geq \left\lceil \beta^k \cdot M^k / r^k_{q,\tau} \right\rceil \qquad \text{, for all } k, \tag{8.14}$$

$$\bar{c}^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}) \geq \alpha^k \qquad \text{, for all } k, \tag{8.15}$$

$$d^k_{q,\tau} \geq \gamma^k \cdot s^k_{q,t}(\boldsymbol{x}) \qquad \text{, for all } k, \boldsymbol{x}, \tag{8.16}$$

$$s^k_{q,\tau}(\boldsymbol{x}) = d^k_{q,\tau} + g^{k,\pi^{up}}_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, \boldsymbol{x}) \qquad \text{, for all } k, \boldsymbol{x}. \tag{8.17}$$

The optimum of (8.12) is found by first finding the solution space for $d^k_{q,\tau}$, $k = 1, \ldots, K$, using constraints (8.13) and (8.14), and the optimal solution of the fixed staffing model, and, second, the solution space for $f_{q,\tau}$ using constraint (8.16). Next, complete enumeration over the obtained solution space is applied, which can be done quickly for realistically sized instances.

Lower bound model. For the lower bound model, we assume that we are allowed to reconsider the nurse-to-care-unit assignment at the start of every time interval. To observe that this relaxation leads to a lower bound on staffing requirements, note that with a given number of nurses, a higher coverage compliance can be achieved than in the original model. The assignment procedure π^{low} is executed at the start of each time interval, and the coverage compliance can thus be calculated per time interval. The coverage compliance over a shift $\underline{c}^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau})$ can then be calcu-

lated by:

$$\underline{c}^k_{q,\tau}\left(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}\right) = \frac{1}{\ell_\tau} \sum_{t=b_\tau}^{b_\tau+\ell_\tau-1} \sum_{\boldsymbol{x}} \left\{ \mathbb{1}\left(x^k \le r^k_{q,\tau} \cdot s^k_{q,t}(\boldsymbol{x})\right) \cdot \prod_{w=1}^{K} \hat{Z}^w_{q,t}(x^w) \right\}.$$

where $s^k_{q,t}(\boldsymbol{x})$ is the number of nurses staffed at care unit k for time interval $[t, t+1)$ on day q under assignment policy π^{low}, when census configuration $\boldsymbol{x}$ is observed at time (q,t).

Since π^{low} is executed every time interval, it is based on the census configuration at the start of that time interval. A nurse from the flex pool gets staffed on the unit where the number of nurses short is the highest:

$$g^{\pi^{low}}_{q,t}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, \boldsymbol{x}) = \max_{\left\{f^1_{q,t},\ldots,f^K_{q,t} : \sum_k f^k_{q,t} = f_{q,\tau}\right\}} \min_k \frac{r^k_{q,\tau} \cdot (d^k_{q,\tau} + f^k_{q,t}) - x^k}{r^k_{q,\tau}}.$$

As a result, for each shift (q,τ), we have:

$$\min z_L = \omega_f f_{q,\tau} + \sum\nolimits_k \omega_d d^k_{q,\tau} \tag{8.18}$$

$$\text{s.t. } d^k_{q,\tau} \ge S^k \qquad \text{, for all } k, \tag{8.19}$$

$$d^k_{q,\tau} \ge \left\lceil \beta^k \cdot M^k / r^k_{q,\tau} \right\rceil \qquad \text{, for all } k, \tag{8.20}$$

$$\underline{c}^k_{q,\tau}\left(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}\right) \ge \alpha^k \qquad \text{, for all } k, \tag{8.21}$$

$$d^k_{q,\tau} \ge \gamma^k \cdot s^k_{q,t}(\boldsymbol{x}) \qquad \text{, } b_\tau \le t < b_\tau + \ell_\tau \text{, for all } k, \boldsymbol{x}, \tag{8.22}$$

$$s^k_{q,t}(\boldsymbol{x}) = d^k_{q,\tau} + g^{k,\pi^{low}}_{q,t}\left(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, \boldsymbol{x}\right) \quad \text{, } b_\tau \le t < b_\tau + \ell_\tau \text{, for all } k, \boldsymbol{x}. \tag{8.23}$$

The optimum of (8.18) is found by first finding the solution space for $d^k_{q,\tau}, k = 1,\ldots,K$, using constraints (8.19) and (8.20), and the optimal solution of the fixed staffing model, and, second, the solution space for $f_{q,\tau}$ using constraint (8.22). Next, complete enumeration over the obtained solution space is applied, which can be done quickly for realistically sized instances.

Flexible staffing levels. The upper and lower bound models were formulated to be able to find, or otherwise approximate, the optimal solution of the flexible staffing model. Here, we discuss how the solutions of the fixed model, and the upper and lower bound models, can be used to select the best staffing configuration. Two questions need to be answered: (1) did we find the optimal solution for the flexible staffing model, and, (2) which staffing configuration to select as the best solution?

Let us first discuss question (1). Observe that $z_L \le z_U$ and $z_L \le z_F$. When $z_L = z_U$ the upper and lower bound coincide so that the optimal solution is found. When $z_L < z_U$, but $z_L = z_F$, the optimal solution is also found, since in this case we are sure that flexible staffing cannot improve upon fixed staffing. In other cases, we

are not sure whether or not the optimal solution is found; then, it is of interest to identify a bound on the distance between the optimal and the obtained solution.

The consideration involved when answering question (2) is to select the solution with the lowest optimal objective value, while it assures that the constraints (8.7)–(8.11) of the flexible staffing model are satisfied. For the solution of the lower bound model we are not sure whether or not constraints (8.7)–(8.11) are satisfied, therefore we never select this solution. In addition, when $z_F = z_U$, as tiebreaker, we choose the solution that achieves the highest minimum coverage compliance.

Let us denote with S_F, S_U, and S_L the optimal staffing configurations in the fixed, upper, and lower bound model respectively. We now provide an overview of the different cases:

(a) $z_L = z_U = z_F$. The optimal solution is found; if $\min_k \overline{c}^k_{q,\tau}(\cdot) \geq \min_k c^k_{q,\tau}(\cdot)$, S_U is selected as the best staffing configuration, otherwise S_F.

(b) $z_L = z_U < z_F$. The optimal solution is found; S_U is selected.

(c) $z_L = z_F < z_U$. The optimal solution is found; S_F is selected.

(d) $z_L < z_F = z_U$. Not sure whether or not the optimal solution is found; if $\min_k \overline{c}^k_{q,\tau}(\cdot) \geq \min_k c^k_{q,\tau}(\cdot)$, S_U is selected, otherwise S_F. The bound on the error margin is $z_U - z_L$.

(e) $z_L < z_U < z_F$. Not sure whether or not the optimal solution is found; S_U is selected; the error bound is $z_U - z_L$.

(f) $z_L < z_F < z_U$. Not sure whether or not the optimal solution is found; S_F is selected; the error bound is $z_F - z_L$.

8.4 Numerical results

This section presents the experimental results. The numerical results in this section are based on the case study as presented in Chapter 7 (Section 7.4). Section 8.4.1 describes additional information on the case study with respect to staffing. Section 8.4.2 validates our approximation approach by investigating the distance between the upper and the lower bound solutions. Finally, Section 8.4.3 illustrates the practical potential of our methodology by returning to a selection of the interventions presented in Chapter 7 and formulating two additional interventions.

8.4.1 Case study description

Recall that the following specialties are taken into account: traumatology (TRA), orthopedics (ORT), plastic surgery (PLA), urology (URO), vascular surgery (VAS), and general surgery (GEN). In the present setting, the patients of the mentioned specialties are admitted to four different inpatient care departments. On floor I, care unit A houses GEN and URO, and unit B VAS and PLA. On floor II, care unit C houses TRA, and unit D ORT.

Working days are divided in three shifts: the day shift (8:00–15:00), the evening shift (15:00–23:00), and the night shift (23:00–8:00). These time intervals do indicate the times that nurses are responsible for direct patient care. Around these time intervals, the working times of the day and evening shift also incorporate time for patient handovers, indirect patient care, and professional development. At all times there should be at least two nurses present at each care unit. According to agreements on working conditions for nurses in all university hospitals in the Netherlands, the contractual number of annual working hours per full time equivalent (FTE) is 1872. The number of hours that one FTE can be employed for direct nursing care, after deduction of time reserved for professional development, holiday hours, and sick leave, is 1525.7 on average (also see [192]). The yearly cost per FTE including all costs and bonuses is roughly €53,000.

The nurse-to-patient ratio targets prescribed by the board of the AMC for the studied care units are 1:4 during the day shifts, 1:6 during the evening shifts, and 1:10 during the night shifts. The current staffing practice is based on the number of beds in service, independent of whether these are occupied or not, and no float nurse pools are employed. Thus, for example, for a care unit size of 24 beds and staffing ratio 1:4, the number of dedicated nurses to staff is always 6. Scarcity of nursing capacity frequently leads to expensive hiring of temporary nurses from external agencies, and to undesirable ad hoc bed closings. Also, the prescribed staffing levels cannot always be realized in practice. As a result, the inpatient care units experience a lack of consistency in the delivered quality of nursing care.

8.4.2 Quality of the bounds

To investigate the performance of the approximation approach for flexible staffing, we test the fixed, the upper, and the lower bound models on a variety of parameter settings. The bed census distributions as were obtained with the prediction model of Chapter 7 for the base case for the year 2010 are taken as input for the three staffing models. Based on the intention of the AMC, we assume that two float nurse pools are created: one serving care units A and B on floor I, and one serving care units C and D on floor II. During the planning horizon of a year, during which no cyclical Master Surgery Schedule (MSS) was used, we thus have to staff $365 \times 3 = 1095$ unique working shifts.

For our set of test instances, Table 8.1 provides an overview of the considered parameter settings. We vary over the (relative) staffing cost for float nurses, the coverage compliance threshold, the minimum coverage requirement, and the minimum dedicated nurse fraction. In addition, three different staffing ratio configurations are considered. We evaluate 2250 instances, together containing 2,463,750 working shifts to be staffed.

For each of the evaluated shifts, we recorded whether the optimum for the flexible staffing model was found or not. Table 8.2 displays the results. The overall result is that in 94.0% of the cases the optimum is found. In addition, the following effects can be observed. The optimum is found more often when flexible staffing is less attractive (which is reflected by increasing β^k and γ^k). Also, the minimum

Table 8.1: Input parameter settings of the test instances for care units $k \in \{A, B, C, D\}$.

Parameter	*Description*	*Value*
Fixed		
Q	Planning horizon in days	365
T	Number of time intervals per day	24
$\lvert\mathcal{T}\rvert$	Number of shift types	3
(b_1, b_2, b_3)	Shift start times	$(8, 15, 23)$
(ℓ_1, ℓ_2, ℓ_3)	Shift durations	$(7, 8, 9)$
S^k	Minimum staffing levels	2
ω_d	Staffing cost dedicated nurse	1
Variable		
ω_f	Staffing cost float nurse	$\{1, 1.25, 1.5\}$
α^k	Minimum coverage compliance	$\{0.75, 0.80, 0.85, 0.90, 0.95\}$
β^k	Minimum coverage	$\{0.5, 0.6, 0.7, 0.8, 0.9\}$
γ^k	Minimum fraction of dedicated nurses	$\{0.5, 0.6, 0.7, 0.8, 0.9\}$
$(r^k_{q,1}, r^k_{q,2}, r^k_{q,3})$	Nurse-to-patient ratio targets	$\{(4, 6, 10), (4, 6, 8), (5, 5, 10)\}$

staffing levels $S^k = 2$ make that for night shifts the fixed and flexible solution generally coincide. Therefore, the optimum is almost always found for these shifts. For decreasing α^k the optimum is found more often, which may seem counterintuitive. However, for lower α^k the minimum coverage requirement given by β^k becomes decisive, which reduces the attractiveness of float nurses.

At the end of Section 8.3.1, we described how to find error bounds on the deviation from the optimal objective value in case one is not sure whether or not the optimum is found. Figure 8.1 zooms in on the 6.0% of shifts for which this holds; it shows a histogram of the deviations per shift of the obtained solution from the lower bound solution. The average maximum deviation for non-optimal shifts is 8.1%. It can be observed that on an individual shift level, the deviation can be substantial, because of the inherent integrality of the number of nurses that can be staffed. By

Table 8.2: The percentage of shifts for which the optimal solution is found (ceteris paribus).

Shift type (τ)		*Float nurse cost* (ω_f)		*Nurse-to-patient ratios* ($r^k_{q,\tau}$)	
day	87.3%	1.00	94.2%	4,6,8	93.8%
evening	94.9%	1.25	93.6%	4,6,10	93.9%
night	99.9%	1.50	94.3%	5,5,10	94.3%
Coverage compliance (α^k)		*Coverage minimum* (β^k)		*Flexibility ratio* (γ^k)	
0.75	96.4%	0.50	82.9%	0.50	91.0%
0.80	95.4%	0.60	89.2%	0.60	91.0%
0.85	94.2%	0.70	98.3%	0.70	91.4%
0.90	93.1%	0.80	99.6%	0.80	96.6%
0.95	90.9%	0.90	100.0%	0.90	100.0%

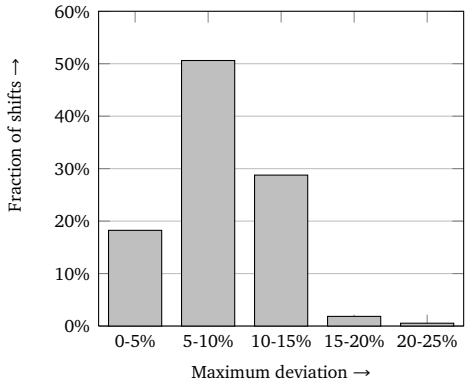


Figure 8.1: Distribution of the distance between the obtained solution and the lower bound solution (non-optimal shifts, $n = 147,426$).

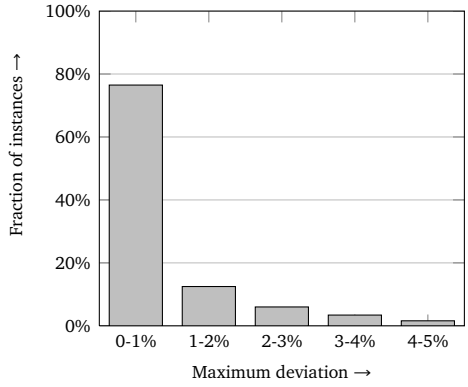


Figure 8.2: Distribution of the error bound on total staffing costs (all instances, $n = 2250$).

displaying the error bound on the total staffing cost per instance, Figure 8.2 shows that the impact of these deviations on the overall performance is small: on average the obtained total staffing costs are within 0.6% of the optimum. We conclude that the approximative approach via bounds on the staffing levels, performs nearly optimal for our case study.

8.4.3 Case study results

To illustrate the potential of the presented staffing methodology for the case study, we return to a selection of the interventions that we presented in Chapter 7, which were formulated to improve the efficiency of the inpatient care service operations in terms of productivity of the inpatient beds.

We investigate both the value of aligning staffing levels with bed census predictions and of employing float nurses, by comparing the results of the fixed and

flexible staffing models with the current staffing policy, which we refer to as 'full staffing'. With M^k the capacity of care unit k in the number of beds, under the full staffing policy always $\lceil M^k / r^k_{q,\tau} \rceil$ nurses are required.

The intended AMC practice will be that registered nurses will alternately be rostered as a dedicated or float nurse. Therefore, we consider the case in which dedicated and float staff members are equally expensive, i.e., $\omega_d = \omega_f$. In addition to the fixed input as displayed in Table 8.1, the board of the AMC has chosen to deploy the following quality of care requirements: nurse-to-patient ratios $r^k_{q,1} = 4$, $r^k_{q,2} = 6$, $r^k_{q,3} = 10$, minimum coverage $\beta^k = 0.70$, coverage compliance $\alpha^k = 0.90$, and at least two out of three nurses should be dedicated nurses, i.e., $\gamma^k = 0.67$. Compared to Chapter 7, we formulate two additional interventions, and we do not consider interventions 2 and 6. For a complete specification of the base case scenario, and intervention 1, 3, 4, and 5, we refer the reader to Chapter 7. The detailed results are presented in Tables 8.3 and 8.4. Table 8.5 provides an overview of the results for the various interventions. It includes the calculation of the following productivity measure: the number of patients treated per employed FTE per year.

Base Case. First, we evaluate the performance of the base case scenario, the situation that most closely resembles current practice. The results are displayed in Table 8.3. In the flexible staffing policy, two flex pools are installed, one on each floor, we therefore present the results per floor. The number of FTEs required is calculated by adding up the total number of staffed nurse hours and dividing by the 1525.7 direct nursing hours that one FTE has available. For the base case we show three values for the coverage compliance threshold ($\alpha^k = \{0.85, 0.90, 0.95\}$), to illustrate the effect of this quality-of-care constraint on required nursing capacity.

For both the fixed and the flexible staffing model, it turns out that the realized coverage compliance is on average much higher than the minimum requirement. This is due to the fact that when the coverage compliance constraint is slightly violated, an additional nurse needs to be staffed, which significantly increases

Table 8.3: The numerical results for the base case (Floor I: 56 beds, 56.7% utilization; Floor II: 48 beds, 58.6% utilization; with the FTE-Δ% relative to full staffing).

Intervention	*Floor*	*Full staffing*	*Fixed staffing*			*Flexible staffing*			
		FTE (#)	Average coverage	FTE (#)	(Δ%)	Error bound (%)	Average coverage	FTE (float) (#)	(Δ%)
Base case									
$\alpha = 0.85$	I	57.7	0.96	44.8	-22.2	0.4	0.96	44.7 (1.7)	-22.4
	II	48.3	0.96	38.9	-19.5	0.0	0.95	38.8 (2.0)	-19.7
$\alpha = 0.90$	I	57.7	0.98	46.0	-20.3	0.8	0.97	45.7 (2.7)	-20.8
	II	48.3	0.97	40.0	-17.3	0.1	0.97	39.6 (2.8)	-18.0
$\alpha = 0.95$	I	57.7	0.99	47.9	-16.9	1.4	0.99	47.4 (4.6)	-17.8
	II	48.3	0.99	42.5	-12.1	0.4	0.99	41.1 (4.3)	-14.9

Table 8.4: The numerical results for the various interventions (with the FTE-Δ% relative to full staffing).

Intervention	*Capacity* (# beds)	*Utilization* (%)	*Full staffing* FTE (#)	*Fixed staffing* Average coverage	FTE (#)	FTE (Δ%)	*Flexible staffing* Average coverage	FTE (float) (#)	FTE (float) (Δ%)
1. Rationalize bed requirements									
Floor I	48	66.1	48.1	0.99	43.8	-8.9	0.98	43.3 (6.2)	-9.9
Floor II	40	70.1	42.6	0.99	39.3	-7.8	0.98	38.7 (5.2)	-9.1
3. Change operational process									
Floor I	45	63.4	48.1	0.98	41.8	-13.0	0.98	41.6 (4.4)	-13.5
Floor II	39	68.3	42.6	0.98	38.4	-9.9	0.98	37.2 (6.9)	-12.7
4. Balance MSS									
Floor I	46	71.3	48.1	0.99	45.7	-5.0	0.99	44.9 (7.8)	-6.7
Floor II	40	71.5	44.5	0.98	40.9	-8.2	0.98	39.6 (6.1)	-11.0
5. Combination (1), (3) and (4)									
Floor I	44	66.9	48.1	0.98	42.4	-11.7	0.98	41.8 (6.4)	-13.1
Floor II	39	69.5	42.6	0.98	38.8	-8.8	0.98	38.1 (4.6)	-10.6
7a. Combination (1) and centralized flex pool									
Floors I & II	88	67.9	90.7	0.99	83.1	-8.4	0.98	80.2 (9.5)	-11.5
7b. Combination (5) and centralized flex pool									
Floors I & II	83	68.1	90.7	0.98	81.3	-10.3	0.98	77.4 (8.6)	-14.6
8a. Combination (7a) and merge care units									
Floors I & II	88	67.9	84.9	0.97	74.7	-12.1	0.96	73.8 (9.7)	-13.1
8b. Combination (7b) and merge care units									
Floors I & II	83	68.1	83.3	0.97	72.0	-13.5	0.97	71.5 (9.6)	-14.1

the coverage compliance since this nurse can care for $r^k_{q,\tau}$ patients. Although full staffing ensures a coverage compliance of 100%, it frequently overstaffs care units. It is clear that the acceptance of slight coverage reductions (still realizing average coverage compliances higher than 95%), allows managers to better match care supply and demand, thereby realizing efficiency gains of 12–22%. The largest gain is achieved by the staffing based on census predictions (see results fixed model). The additional value of employing float nurses is case dependent, and in most cases higher with increasing α^k, due to the increasing gap with the minimum coverage requirement set by β^k.

Interventions 1,3,4,5. Intervention 1 rationalized the care unit dimensions based on the requirement of rejection probabilities not exceeding 1%, 2.5%, and 5%. We focus on the outcomes for 2.5%; this is the threshold selected by the AMC to be implemented in practice. Table 8.4 shows that fixed staffing with $\alpha^k = 0.9$ reduces nursing capacity requirements by 8–9% compared to full staffing, and flexible staffing yields an additional 1% reduction. Table 8.5 indicates the gain

Table 8.5: FTE and productivity results for all interventions (with both the FTE-Δ% and the productivity-Δ% relative to full staffing in the base case).

Intervention	*Full staffing*				*Fixed staffing*				*Flexible staffing*			
	FTE		Productivity		FTE		Productivity		FTE		Productivity	
	(#)	(Δ%)	(#/yr)	(Δ%)	(#)	(Δ%)	(#/yr)	(Δ%)	(#)	(Δ%)	(#/yr)	(Δ%)
Base case	106.0	-	42.3	-	85.9	-18.9	52.2	+23.3	85.3	-19.5	52.6	+24.2
(1)	90.7	-14.4	48.5	+14.5	83.1	-21.6	52.9	+25.0	82.1	-22.6	53.5	+26.5
(3)	90.7	-14.4	48.4	+14.4	80.2	-24.3	54.7	+29.4	78.7	-25.7	55.8	+31.8
(4)	92.6	-12.6	48.6	+14.8	86.5	-18.4	52.0	+22.8	84.5	-20.3	53.2	+25.8
(5)	90.7	-14.4	49.6	+17.2	81.3	-23.3	55.3	+30.7	79.8	-24.7	56.3	+33.0
(7a)	90.7	-14.4	48.5	+14.5	83.1	-21.6	52.9	+25.0	80.2	-24.3	54.8	+29.5
(7b)	90.7	-14.4	49.6	+17.2	81.3	-23.3	55.3	+30.7	77.4	-27.0	58.1	+37.2
(8a)	84.9	-19.9	51.7	+22.3	74.7	-29.5	58.8	+39.0	73.8	-30.3	59.5	+40.7
(8b)	83.3	-21.4	54.0	+27.6	72.0	-32.0	62.4	+47.5	71.5	-32.5	62.8	+48.5

Productivity: number of patients treated per employed FTE per year

against current practice: 22.6% reduction in FTE requirements, with a simultaneous increase of staff productivity by 26.5%.

Intervention 3 focused on changes in the operational process by: (a) decreasing lengths of stay by admitting all elective patients on the day of surgery, and (b) reducing afternoon census peaks by encouraging discharges to take place before noon. The reduction of demand and its variability lowered the number of beds required. Here we see that our staffing methodology also translates this into significantly lower staff requirements, and higher productivity.

Intervention 4 intended to decrease artificial demand variability by designing a cyclical Master Surgery Schedule (MSS) with the purpose to balance bed census. Recall that due to the integrality of the number of scheduled operating room blocks, the resulting MSS slightly increased patient demand. Therefore, its impact on staffing requirements is not directly visible. However, its impact is revealed by the outcomes on the fifth intervention (the combination between interventions 1, 3, and 4) which outperforms all previous configurations on the productivity measure.

Finally, let us state two general insights. First, note that under the old (full) staffing policy, a reduction in the number of beds not always translates into a reduction in staffing requirements. This is the case when the number of beds does not decrease to a capacity level such that it crosses a level that is a multiple of one of the nurse-to-patient ratios. Second, from our results we cannot deduce general rules-of-thumb for the potential of float nurses. The outcomes for each particular care unit are a complex interplay between care unit sizes, nurse-to-patient ratios, and the shapes of the bed census distributions.

Interventions 7 and 8. The first additional intervention involves the merging of the two flex pools into one flex pool which serves all four care units. Intervention 7a

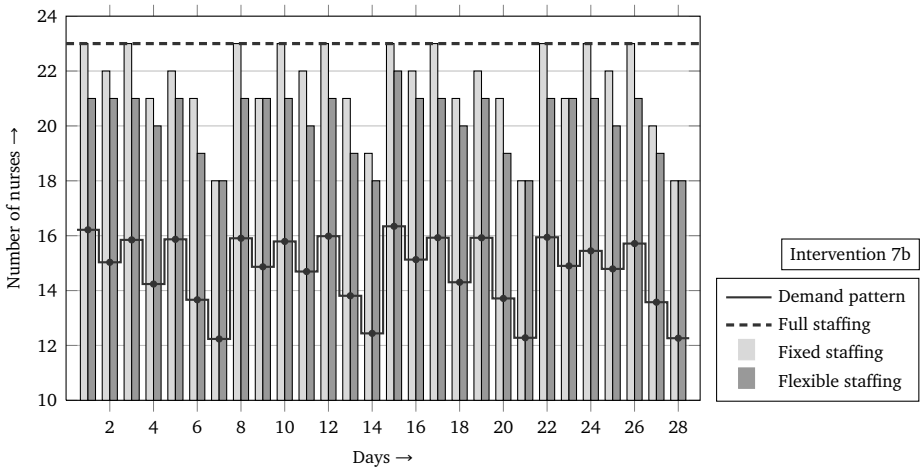


Figure 8.3: Staffing levels for day shifts on Floor I during the 4-week period starting on Monday January 25 (the demand pattern shows the average census divided by ratios $r^k_{q,1}$).

evaluates the impact of this centralized flex pool for the situation of intervention 1, and intervention 7b for that of intervention 5. Naturally, for the full and fixed staffing policies the outcomes for intervention 7a and 7b coincide with 1 and 5 respectively, due to the unchanged care unit sizes and bed census distributions. With the flexible staffing policy, the additional flexibility of having four instead of two allocation options for each float nurse pays off: an additional saving of around 1.5–2.5 FTEs can be realized, in conjunction with an additional productivity increase of 3–4%. As an illustration, for intervention 7b, the effect of staffing levels following bed census demand patterns and the difference between fixed and flexible staffing therein are visualized in Figure 8.3.

Intervention 8 merges care units A and B, and care units C and D (intervention 8a for the situation of intervention 1, and intervention 8b for that of intervention 5). The two remaining care units, floor I and floor II, share one flex pool. This intervention is hard to implement, because it would imply the necessity of thorough renovation of the building. Although fictitious on the short-term, the positive outcomes for this intervention show that it is worthwhile considering. The economies-of-scale effect shows itself in various ways. First, larger care unit sizes reduce the occurrence of overstaffing due to staffing levels that have to be rounded upwards as a result of the nurse-to-patient ratios. Second, the relative variation in bed census decreases, making it easier to align staffing levels with patient demand, which is expressed by the results for the fixed staffing model. Third, in this case the minimum staffing levels of $S^k = 2$ per care unit only need to be satisfied for two care units, which often results in decreased staffing requirements during night shifts. Finally, it can be observed that the additional value of employing float nurses is lower for larger care unit sizes, again due to the decreasing relative census variation.

8.5 Discussion

Rising healthcare costs and increasing nurse shortages make cost-effective nurse staffing of utmost importance. In many hospitals, staffing levels are a result of historical development, as hospital managers lack the tools to base staffing decisions on information about future patient demand. Since patient safety is jeopardized when medical care units are understaffed, scarcity of nursing capacity can lead to expensive hiring of nurses from external agencies and to undesirable ad hoc bed closings. In this chapter, we have presented a generic analytical method that can quantitatively support decision making on required staffing levels in inpatient care facilities. We have demonstrated its potential with a case study of the AMC, for which we have shown that by achieving coherence between patient demand and staffing supply simultaneous cost reductions and quality of care improvements are possible.

The combined application of the bed census prediction model from Chapter 7 and the staffing models from the current chapter enables hospital administrators to gain insight into the value of integrated decision making. The interrelation between decisions such as case mix, care unit partitioning, care unit size, and admission/discharge times is made explicit. Because the demand prediction model incorporates the operating room block schedule and the patient arrival pattern from the emergency department, the presented methodology also facilitates alignment between the design and operations of the inpatient care facility and its surrounding departments. With this integrated framework, staffing effectiveness can be attained in three steps. First, the method can help to reduce artificial variability of bed occupancies, for example by adjusting the operating room schedule. Second, by predicting the bed census distributions and determining staffing levels for dedicated nurses accordingly, the predictive part of the remaining variability can be anticipated. Third, to be able to effectively respond to random variability, adequately sized float nurse pools can be created.

Staffing requirements are the result of a complex interaction between care unit sizes, nurse-to-patient ratios, the bed census distributions, and the quality-of-care requirements. The optimal configuration strongly depends on the particular characteristics of a specific case under study. Nonetheless, several general insights have been obtained. When working with nurse-patient-ratios, care units should be sufficiently large, to avoid efficiency losses due to the lack of granularity in the values of the ratios. Next, under the premise that the costs per float nurse remain unchanged, the more care units float nurse pools can serve, the more effective they are. Finally, also when it does not reduce capacity requirements, flexible staffing staffing is beneficial since it enhances the adherence to the nurse-to-patient ratio targets.

The case study of the AMC provides an example of how the methodology can be applied in practice. Due to both economic and medical developments, the AMC is forced to reorganize the operations of the inpatient services during the upcoming years. Nurse staffing is high on the agenda, since the AMC has 30 inpatient departments, staffing costs account for 66% of the total expenses in the AMC, and

one full-time registered nurse yearly costs around €53,000. We have applied our staffing models on data of several care units; for four of them we presented the results in this chapter. The formulations of all interventions and the eventual parameter settings are the results of close cooperation between operations researchers and hospital managers from different levels within the organization. It has resulted in the joint conclusion that efficiency gains are possible, while improving upon the adherence to nurse-to-patient ratio targets. As a result, the AMC decided that the flexible nurse staffing method will be fully implemented during the upcoming years.

The development of a user-friendly decision support system (DSS) based on our method will be a next step in achieving practical impact. Our model relies on data that is easily extractable from typical hospital management systems. This makes it possible to automate the process of collecting the required input parameters to run the model. Integration with the hospital management system, visualization of the results, and the possibility to run what-if scenarios will be desired specifications of the DSS. We believe that the adoption of such a system by healthcare administrators of inpatient care services can result in more cost-effective resource capacity planning and control decisions.

8.6 Appendix

8.6.1 Complexity of the flexible staffing model

This appendix investigates the complexity of the calculations involved in solving the flexible staffing model, formulated by equations (8.6)–(8.11). The complexity is such that the computation time inhibits the evaluation of realistically sized instances. This is mainly due to the large number of census configurations that has to be evaluated to identify the float nurse assignment procedure π^* satisfying the maximization (8.5). This assignment procedure is involved in constraint (8.11).

Consider shift (q,τ). Let us investigate the complexity of determining π^* for a given availability of dedicated and float nurses, i.e, for given $d^1_{q,\tau},\ldots,d^K_{q,\tau},f_{q,t}$. For every census configuration $\boldsymbol{y}$ that can possibly be observed at the start of the shift, the assignment (8.5), to be used in (8.11), needs to be found. This is of order:

$$O(N_y \cdot N_f \cdot N_c \cdot K),$$

where N_y denotes the maximum number of possible census configurations at the start of the shift, N_f the number of possible allocations of the $f_{q,\tau}$ available float nurses, and N_c the complexity of the calculations involved in evaluating the coverage compliance $c^k_{q,\tau}(\boldsymbol{d}_{q,\tau},f_{q,\tau},r^k_{q,\tau};\boldsymbol{y})$, which has to be done for all K wards.

Since the census range for ward k is $\{0,\ldots,M^k\}$, with $\hat{M}=\max_k M^k$, we have:

$$N_y=(\hat{M}+1)^K.$$

Second, counting the number of possible allocations of $f_{q,\tau}$ nurses over K wards, we

have:

$$N_f = \binom{f_{q,\tau} + K - 1}{K - 1}$$

This leaves us to determine N_c. To this end, we make use of the concept patient *cohort* (as also introduced in Chapter 7): a cohort is a group of patients originating from a single instance of an OR block (electives) or admission time interval (acute patients). As specified in Chapter 7, all patients of one cohort are preferably placed on the same care unit. The best coverage is realized when for each patient cohort at the start of the shift it is observed how many patients are present, since in that manner the maximum amount of information on possible admissions and discharges is taken into consideration. Let Φ denote the total number of patient cohorts present during shift (q, τ), and $\mathcal{W}^k$ the set of patient cohorts admitted to ward k. For notational convenience we introduce the function $v^i_{q,t}$ as $v^i_{q,t} = h^i_{q,t}$ for the elective patients, and $v^i_{q,t} = g^i_{q,t}$ for acute patient types. In addition, for each patient cohort, we define for $b_\tau \leq t < b_\tau + \ell_\tau$ the conditional distribution $v^{i,z_i}_{q,t}$, with $v^{i,z_i}_{q,t}(x_i)$ the probability that x_i patients of cohort i are present at the start of time interval (q, t), given that at the start of shift (q, τ) the number of patients present of this cohort was z_i. Then, the coverage compliance given that census configuration $\boldsymbol{y}$ is observed at the start of shift (q, τ) is:

$$c^k_{q,\tau}(\boldsymbol{d}_{q,\tau}, f_{q,\tau}, r^k_{q,\tau}; \boldsymbol{y}) = \sum_{\substack{z_1,\ldots,z_\Phi: \\ \sum_{i\in\mathcal{W}^k} z_i = y^k, \\ k=1,\ldots,K}} \left\{ \prod_{i=1}^{\Phi} v^i_{q,b_\tau}(z_i) \cdot \frac{1}{\ell_\tau} \sum_{t=b_\tau}^{b_\tau+\ell_\tau+1} \sum_{x_i:\forall i\in\mathcal{W}^{\|}} \mathbb{1}\Big(\sum_{i\in\mathcal{W}^k} x_i \leq r^k_{q,\tau} \cdot s^k_{q,\tau}(\boldsymbol{y}) \Big) \Big\{ \prod_i v^{i,z_i}_{q,t}(x_i) \Big\} \right\}.$$

The first summation involves maximally $(\hat{M}+1)^\Phi$ combinations, the second summation ℓ_τ combinations, and the third $\hat{M}+1$. Therefore, we have

$$N_c = (\hat{M}+1)^{\Phi+1} \cdot \ell_\tau.$$

To conclude, the complexity of determining π^* for given $d^1_{q,\tau}, \ldots, d^K_{q,\tau}, f_{q,t}$ is of the order:

$$O(N_{\boldsymbol{y}} \cdot N_f \cdot N_c \cdot K) = O\left(\left(\hat{M}+1\right)^{K+\Phi+1} \cdot \binom{f_{q,\tau} + K - 1}{K - 1} \cdot \ell_\tau \right),$$

which for real-world instances is both in terms of memory and computation time too large to find the optimal $d^1_{q,\tau}, \ldots, d^K_{q,\tau}, f_{q,t}$.

8.6.2 Derivation maximum census

In this appendix, $\hat{W}^k_{q,\tau}$ is derived, the probability distribution of the maximum census at care unit k during shift (q,τ). For each patient cohort and each shift (q,τ), we need to determine at which of the time points $t \in \{(q,b_\tau),\ldots,(q,b_\tau+\ell_\tau-1)\}$ the number of patients of this cohort reaches its maximum.

We first determine for each cohort i, the probability distribution $w^i_{q,\tau}$ for the maximum number of patients of this cohort present during shift (q,τ). Since all patients of one cohort are preferably placed on the same care unit, to obtain the probability distribution $W^k_{q,\tau}$ for the maximum demand for unit k during shift (q,τ), we take the discrete convolution over the distributions $w^i_{q,\tau}$ relevant to unit k. Finally, from the maximum demand distribution $W^k_{q,\tau}$, the maximum census distribution $\hat{W}^k_{q,\tau}$ is obtained by applying the same transformation as was done for $Z^k_{q,\tau}$ and $\hat{Z}^k_{q,\tau}$ in Chapter 7 in equation (7.1).

Elective patients. For each combination of a day q in the Inpatient Facility Cycle (IFC, see Chapter 7), and a number of days after surgery n, there is a unique corresponding day in the Master Surgery Schedule (MSS, see Chapter 7). We denote this day by $\Delta^{MSS}(q,n)$:

$$\Delta^{MSS}(q,n) = \begin{cases} (q-n) \bmod S + \mathbb{1}_{\left((q-n) \bmod S=0\right)} \cdot S & ,-1 \le n < q, \\ (q-n) + \left[((n-q) \operatorname{div} S)+1\right] \cdot S & ,q \le n \le L^i. \end{cases}$$

Also, note that by definition of the cohorts, the combination of day q and cohort i uniquely defines the number of days the patients of this cohort are already present after surgery; let us denote this value by $N(i,q)$. For elective patients, $w^i_{q,\tau}$ can be calculated as follows. For all i such that $\exists i$ such that $i \in b_{i,\Delta^{MSS}(q,N(i,q))}$:

$$w^i_{q,\tau} = \begin{cases} h^i_{N(i,q),b_\tau} & ,N(i,q) = 1,\ldots,L^i, \\ h^i_{0,b_\tau} & ,N(i,q) = 0, \vartheta_i < b_\tau, \\ h^i_{0,\vartheta_i} & ,N(i,q) = 0, b_\tau \le \vartheta_i < b_\tau+\ell_\tau, \\ h^i_{0,b_\tau+\ell_\tau-1} & ,N(i,q) = 0, \vartheta_i \ge b_\tau+\ell_\tau, \\ h^i_{-1,b_\tau+\ell_\tau-1} & ,N(i,q) = -1, b_\tau+\ell_\tau \le T, \\ h^i_{-1,T+\vartheta_i} & ,N(i,q) = -1, b_\tau+\ell_\tau > T, \vartheta_i < b_\tau+\ell_\tau-T, \\ h^i_{-1,b_\tau+\ell_\tau-1} & ,N(i,q) = -1, b_\tau+\ell_\tau > T, \vartheta_i \ge b_\tau+\ell_\tau-T. \end{cases}$$

Acute patients. Let $\Delta^{AAC}(q,n)$ be the admission day in the Acute Admission Cycle (AAC, see Chapter 7) of an acute patient type present on a given day q in the IFC, and which is at its n-th day after admission:

$$\Delta^{AAC}(q,n) = \begin{cases} (q-n) \bmod R + \mathbb{1}_{\left((q-n) \bmod R=0\right)} \cdot R & ,0 \le n < q, \\ (q-n) + \left[((n-q) \operatorname{div} R)+1\right] \cdot R & ,q \le n \le L^i. \end{cases}$$

Recall that an acute patient type is identified by (p, r, θ). Observe that an acute patient cohort i is specified by the combination of a patient type j and a specific admission day. Also for acute patients, the combination of day q and cohort i uniquely defines the number of days the patients of this cohort are already present; let us denote this value by $M(i,q)$. During shift (q,τ), for an acute patient cohort the maximum demand is obtained at its admission time interval if this lies within (q,τ), otherwise it is obtained at the start of the shift. Hence, for acute patients $w^i_{q,\tau}$ is calculated by:

$$w^i_{q,\tau} = \begin{cases} g^i_{M(i,q),b_\tau} & ,M(i,q) = 1,\dots,L^i, i \text{ such that } \Delta^{AAC}(q,M(i,q)) = r, \\ g^i_{0,b_\tau} & ,M(i,q) = 0, \theta < b_\tau, i \text{ such that } \Delta^{AAC}(q,M(i,q)) = r, \\ g^i_{0,\theta} & ,M(i,q) = 0, b_\tau \le \theta < b_\tau + \ell_\tau, i \text{ such that} \\ & \Delta^{AAC}(q,M(i,q)) = r, \\ g^i_{0,\theta} & ,M(i,q) = 0, b_\tau + \ell_\tau > T, \theta < b_\tau + \ell_\tau - T, i \text{ such that} \\ & \Delta^{AAC}\left((q+1) \bmod Q + Q \cdot \mathbb{1}_{((q+1) \bmod Q=0)}, M(i,q)\right) = r. \end{cases}$$

Finally, by taking the discrete convolution over the distributions $w^i_{q,\tau}$ relevant to unit k, distribution $W^k_{q,\tau}$, $k = 1,\dots,K$ is obtained. Then, distribution $\hat{W}^k_{q,\tau}$, $k = 1,\dots,K$, is obtained by applying the transformation as presented in equation (7.1).

Part VI

Modeling Care Chains with Stochastic Petri Nets

CHAPTER 9

Introduction

9.1 Motivation

Healthcare organizations typically consist of many departments and serve a wide variety of patient types. Pathways of patients are generally stochastic and various patient flows share different resources. Typical questions arising are identification of bottlenecks, achievable throughput and maximization of resource utilization. Therefore, performance analysis is an important issue in the design and implementation of healthcare systems. Below, we argue that an appropriate formalism to model interacting care pathways in healthcare organizations is that of 'stochastic Petri nets'. In the upcoming chapters, we establish a stepping stone for a theoretical framework along which vital insight in the behavior of healthcare networks can be obtained.

Competition over resources is an important issue in many practical systems. Besides healthcare environments, examples of such systems are computer systems, telecommunication networks, flexible manufacturing systems. Several approaches exist for performance analysis of complex systems, such as discrete-event simulation, numerical approximations or exact analytical results. Obtaining analytical results has two main advantages. First, it provides vital insight in the qualitative behavior of involved systems, so that the key characteristics of a system can be detected. In particular, qualitative results related to the structure of the system are often of great importance. Second, it enables efficient computation of relevant performance measures. In many theoretical and practical studies of performance models involving stochastic effects, the statistical distribution of items (customers, jobs, etc.) over places (workstations, queues, etc.) is of great interest, since various of performance measures can be computed from this distribution.

Three main formalisms exist for obtaining analytical closed form results for networks: queueing networks, stochastic process algebras and stochastic Petri nets. The selection of a specific formalism when studying a system preferably depends on the characteristics under investigation. Queueing networks are most suitable when the queueing structure at different locations in the network is the key aspect of the system. When a system consists of building blocks of different processes that are composed into a network, stochastic process algebras may be preferred. Stochastic Petri nets are appropriate when the flow of items and information through the network is the main feature of the system. Since we are interested in the interaction

of flows occurring within healthcare environments, we will focus on the formalism of stochastic Petri nets. When a specific formalism is applied, all network characteristics and all results are preferably formulated in the semantics of that formalism. Therefore, all results are formulated in terms of the Petri net structure, and mainly given in terms of *P*- and *T*-invariants, the central concepts in Petri Nets.

Composition and decomposition of closed form results contribute to less computational effort requirements and greater understanding of network behavior and performance. They allow for studying a system by analyzing the characteristics of separate components. In the following chapters, we study closed form results for the equilibrium distribution of the number of tokens at the places of a stochastic Petri net and the decomposition of this equilibrium distribution into several components corresponding to subnets of the stochastic Petri net.

One of the most important analytical results for the equilibrium distribution describing the number of items at places in a performance model is the so-called *product form* equilibrium distribution found for a fairly wide class of theoretical queueing models. However, practical performance models seldom satisfy the product form conditions. Still, results obtained via the theoretical product form distributions are used for practical problems since these results are found to be robust, that is models which violate the product form conditions are often found to behave in a way very similar to a product form counterpart. The obvious advantages of these product form distributions are their simplicity, since the network behavior is captured in closed form in only a limited set of parameters. This makes product form solutions easy and powerful to use for computational reasons as well as for theoretical reflections for performance models involving congestion. Another important advantage of product form solutions is that it enables us to break down the analysis of a network in the analysis of separate components of the network.

Acting upon the above motivation, the topics of Chapters 10–12 are product form and decomposition for stochastic Petri nets. The research described in these chapters is only a starting point in realizing actual practical healthcare modeling and decision support. Therefore, Chapter 13 formulates suggestions to direct future research.

In the current chapter, in Section 9.2 we first give a detailed description of our contributions. Section 9.3 provides a thorough introduction into the (stochastic) Petri net formalism. Concluding this introductory chapter, Section 9.4 provides a detailed literature survey of product form results and decomposition.

9.2 Contributions

A form of *local balance* is a common element for most performance models with a product form equilibrium distribution. In Chapter 10, *group-local-balance* will be shown to be the concept identifying that the equilibrium distribution of a stochastic Petri net is of product-form nature. Boucherie and Van Dijk [69] presented the group-local-balance concept as the basis for the analysis of batch routing queueing networks. Chapter 10 provides a translation of these results into Petri net terminology. The results on the Markov chain level then provide the foundation to discuss and further investigate structural Petri net implications. We survey the various struc-

tural results that are known for stochastic Petri nets with a product form equilibrium distribution over the number of tokens at the places [66, 68, 129, 182, 270, 296, 406]. The product form results for stochastic Petri nets known from the literature will shown to be unified by group-local-balance, as it forms the connecting principle between these results and the results known for batch routing queueing networks [69, 299]. The results are derived and presented step-by-step to provide an intuitive understanding of the Petri net structure underlying the product form results.

The first structural product form results for stochastic Petri nets were presented by Henderson et. al. [296]. These results are based on the assumption that a positive solution exists for a linear set of equations similar to the traffic equations for queueing networks. It will be shown that group-local-balance implies a positive solution to this linear set of equations, known as the *routing chain*, to exist. A characterization of the structure of the Petri net that is necessary and sufficient for the existence of a positive solution to the routing chain was provided by Boucherie and Sereno [66]. We show that this characterization implies that group-local-balance requires the stochastic Petri net to be an $\mathcal{S}\Pi$-net [270], a stochastic Petri net in which each transition is covered by a minimal support T-invariant. Taking group-local-balance as a starting point enables us to provide additional structural implications and a more intuitive explanation of the known results. By formulating every result in terms of the Petri net structure given by the T-invariants, we also provide structural insights for results known at an algebraic level.

In Chapter 11, from the detailed understanding of the structure behind product form results, we are able to establish a decomposition result. This decomposition result is a generalization of the results obtained by Frosch and Natarajan [222, 223] for closed synchronized systems of stochastic sequential processes, a class of Petri nets in which state machines are synchronized via buffer places. The decomposition result is completely formulated in terms of P- and T-invariants. Similar to buffer places, we define conflict places, which are places that are shared by different minimal closed support T-invariants. Using the P-invariants to assign conflict places as surplus places, places that can be omitted in characterizing the marking of the Petri net, we obtain an algorithmic procedure to verify whether product form holds and for decomposition of the stochastic Petri net into subnets. These subnets correspond to one or more common input bag classes, equivalence classes of T-invariants of the stochastic Petri nets that share an input bag.

Chapter 12 takes the results from Chapter 11 as starting point to formulate an additional decomposition result. It focuses on the subclass of $\mathcal{S}\Pi$-nets that have a product form equilibrium distribution irrespective of values of the transition rates. These nets where algebraically characterized by Haddad et al. [270]. By providing an intuitive interpretation of this algebraical characterization, and associating a state machine to each of the common input bag classes, we obtain a one-to-one correspondence between the marking of the original places and the places of the added state machines. This enables us to show that this subclass of $\mathcal{S}\Pi$-nets can be decomposed into subnets which separate all the common input bag classes of the original net.

The results from Chapters 10–12 form a theoretical foundation to come to performance evaluation of healthcare systems via the formalism of stochastic Petri nets. Chapter 13 provides an outline for future work, which will include deriving approximating results for stochastic Petri nets that do not have a product form equilibrium distribution, and constructing and evaluating stochastic Petri nets based on event logs that can be extracted from electronic database systems of healthcare organizations.

Summarizing, our contributions are the following:

1. We survey the various structural results that are known for stochastic Petri nets with a product form equilibrium distribution over the number of tokens at the places and rephrases all these results in terms of *T*-invariants (Chapter 10).

2. We unify and extend the product form results for stochastic Petri nets by showing that *group-local-balance* can be identified as the concept underlying all these structural results and we provide additional structural implications and an intuitive explanation of the known and new results, all based on *T*-invariants only (Chapter 10).

3. We provide a decomposition result that is completely formulated in terms of both *P*- and *T*-invariants and their derivatives as will defined: common input bag classes, conflict places and surplus places (Chapter 11).

4. We provide an interpretation of the algebraic characterization by [270] of stochastic Petri nets that have a product form equilibrium distribution irrespective of the values of the transition rates. This is accomplished by adding 'bag count places' to the original that form state machines which describe the marking of the original places of a Petri net (Chapter 12).

5. By combining contributions 3. and 4., an additional decomposition result is presented which shows that stochastic Petri nets that have a product form equilibrium distribution irrespective of the values of the transition rates can be decomposed in all their common input bag classes (Chapter 12).

Taking these contributions as a starting point, and with the intention to realize a theoretical framework by which performance evaluation of complex healthcare systems can be achieved via the formalism of stochastic Petri nets, we provide promising directions for future research (Chapter 13).

9.3 Preliminaries

The aim of this section is to provide a general introduction into the formal Petri net language and the Petri net concepts that will be relevant for the analysis in subsequent sections. First, basic definitions of Petri nets and stochastic Petri nets are presented. Next, structural and behavioral properties are introduced. Also, some results derived from these properties of a Petri net that will be used in subsequent sections are listed.

9.3.1 Petri nets

Definitions, properties and results will be presented schematically to provide the reader a convenient reference to the numerous concepts. More elaborate overviews of definitions, properties and results can be found in the survey of Murata [448] and the book of Peterson [480].

Definitions

Definition 9.1 (**Petri net**)**.** A Petri net is a weighted bipartite graph with nodes being either places or transitions and is defined by the 4-tuple $\mathcal{PN} = (P, T, I, O)$, where

- $P = \{p_1, \ldots, p_N\}$ is a finite set of places,
- $T = \{t_1, \ldots, t_M\}$ is a finite set of transitions,
- $I, O : P \times T \to \mathbb{N}$ are the input and output functions identifying the relation between the places and the transitions.

Definition 9.2 (**Marking**)**.** A *marking* $\boldsymbol{m} = (m(n),\ n = 1, \ldots, N)$ of a Petri net is a vector in $\mathbb{N}_0^N$, where $m(n)$ represents the number of *tokens* at place p_n.

Definition 9.3 (**Marked Petri net**)**.** A marked Petri net is a Petri net defined by the 5-tuple $(\mathcal{PN}, \boldsymbol{m}_0) = (P, T, I, O, \boldsymbol{m}_0)$, where $\boldsymbol{m}_0$ is the initial marking.

Definition 9.4 (**Input bag - Output bag**)**.** $I(\cdot,\cdot)$ and $O(\cdot,\cdot)$ give the vectors $\boldsymbol{I}(t) = (I_1(t), \ldots, I_N(t))$ and $\boldsymbol{O}(t) = (O_1(t), \ldots, O_N(t))$, where $I_n(t) = I(p_n, t)$, and $O_n(t) = O(p_n, t)$. The vectors $\boldsymbol{I}(t)$ and $\boldsymbol{O}(t)$ are called the *input* and *output bags* of transition $t \in T$, respectively representing the number of tokens required at the places to fire transition t, and the number of tokens released to the places after firing transition t.

Definition 9.5 (**Transition enabling and firing**)**.** A necessary and sufficient condition for transition t to be *enabled* in marking $\boldsymbol{m}$ is that $m(n) \geq I_n(t)$. When transition t *fires*, then the next state of the Petri net is $\boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{I}(t) + \boldsymbol{O}(t)$. Symbolically this is denoted as $\boldsymbol{m}[t > \boldsymbol{m}'$.

Definition 9.6 (**Firing sequence**)**.** A finite sequence of transitions $\sigma = t_{\sigma_1} t_{\sigma_2} \cdots t_{\sigma_k}$ is a finite *firing sequence* of the Petri net if there exists a sequence of markings $\boldsymbol{m} = \boldsymbol{m}_{\sigma_1}, \ldots, \boldsymbol{m}_{\sigma_{k+1}} = \boldsymbol{m}'$ for which $\boldsymbol{m}_{\sigma_i}[t_{\sigma_i} > \boldsymbol{m}_{\sigma_{i+1}},\ i = 1, \ldots, k$. Symbolically this will be denoted as $\boldsymbol{m}[\sigma > \boldsymbol{m}'$.

Definition 9.7 (**Incidence matrix**)**.** The *incidence matrix* $\boldsymbol{A}$ with entries $A(p,t) = O_p(t) - I_p(t)$ describes the change in the number of tokens in place p when transition t fires, $p \in P, t \in T$.

Definition 9.8 (**Firing count vector**)**.** A vector $\bar{\boldsymbol{\sigma}}$ is the *firing count vector* of the firing sequence σ if $\bar{\sigma}(t)$ equals the number of times transition t occurs in the firing sequence σ.

Definition 9.9 (**State equation**)**.** If $\boldsymbol{m}_0[\sigma > \boldsymbol{m}$, then $\boldsymbol{m} = \boldsymbol{m}_0 + \boldsymbol{A}\bar{\sigma}$. This equation is referred to as the *state equation* for the Petri net.

Definition 9.10 (**Closed set**)**.** For $\mathcal{T} \subseteq T$ define $\mathcal{R}(\mathcal{T})$, the set of input and output bags for the transitions in $\mathcal{T}$, as $\mathcal{R}(\mathcal{T}) = \bigcup_{t\in\mathcal{T}}\{\boldsymbol{I}(t) \cup \boldsymbol{O}(t)\}$. $\mathcal{R}(\mathcal{T})$ is a closed set if for all $\boldsymbol{r} \in \mathcal{R}(\mathcal{T})$ there exist $t, t' \in \mathcal{T}$ such that $\boldsymbol{r} = \boldsymbol{I}(t)$, as well as $\boldsymbol{r} = \boldsymbol{O}(t')$, that is if each output bag is also an input bag, and each input bag is also an output bag for a transition in $\mathcal{T}$.

Definition 9.11 (**State machine**)**.** A Petri net $\mathcal{PN}$ is a *state machine* if and only if $\sum_p I_p(t) = 1$ and $\sum_p O_p(t) = 1$ for all transitions.

Properties

Two types of properties are distinguished. Properties which depend on the initial marking are referred to as *behavioral* and those which are independent on the initial marking as *structural*. Behavioral and structural properties will respectively be marked by the labels [B] and [S].

Definition 9.12 (**Reachability [B]**)**.** A marking $\boldsymbol{m}'$ is *reachable* from marking $\boldsymbol{m}_0$ if a firing sequence σ exists such that $\boldsymbol{m}_0[\sigma > \boldsymbol{m}'$.

Definition 9.13 (**Reachability set [B]**)**.** The *reachability set* $\mathcal{M}(\mathcal{PN}, \boldsymbol{m}_0)$ is a subset of $\mathbb{N}^N$ and gives all reachable markings of the Petri net with initial making $\boldsymbol{m}_0$.

Definition 9.14 (***T*-invariant [S]**)**.** A vector $\boldsymbol{x} \in \mathbb{N}_0^M$ is a *T-invariant* if $\boldsymbol{x} \neq 0$, and $\boldsymbol{Ax} = 0$. From the state equation we obtain that a *T*-invariant represents a firing sequence that brings a marking back to itself (Murata [448]). So *T*-invariants define potential cycles in the reachability set.

Definition 9.15 (***P*-invariant [S]**)**.** A vector $\boldsymbol{y} \in \mathbb{N}_0^N$ is a *P-invariant* (sometimes called *S*-invariant) if $\boldsymbol{y} \neq 0$, and $\boldsymbol{yA} = 0$. *P*-invariants correspond to the conservation of tokens in subsets of places. A *P*-invariant identifies a set of places such that the weighted sum of the number of tokens distributed over these places remains constant for all markings in the reachability set.

Definition 9.16 (**Support [S]**)**.** The *support* of a *T*-invariant $\boldsymbol{x}$ or *P*-invariant $\boldsymbol{y}$ is the set of transitions or places respectively corresponding to non-zero entries of $\boldsymbol{x}$ and $\boldsymbol{y}$, and are denoted by $\|\boldsymbol{x}\|$ and $\|\boldsymbol{y}\|$, i.e., $\|\boldsymbol{x}\| = \{t \in \mathcal{T} \mid x(t) > 0\}$ and $\|\boldsymbol{y}\| = \{p \in \mathcal{P} \mid y(p) > 0\}$.

Definitions 9.17 and 9.18 are stated in terms of *T*-invariants. The definitions are analogous for *P*-invariants.

Definition 9.17 (**Minimal invariant [S]**)**.** A *T*-invariant is a *minimal T-invariant* if there is no other *T*-invariant $\boldsymbol{x}'$ such that $x'(t) \leq x(t)$ for all t.

Definition 9.18 (**Minimal support invariant [S]**)**.** The support of an invariant is minimal if no proper nonempty subset of the support is also the support of a *T*-invariant. An invariant with minimal support is a *minimal support invariant*.

Definition 9.19 (**Closed *T*-invariant [S]**)**.** A *T*-invariant is *closed* if the set of input and output bags for the transitions in its support, $\mathcal{R}(\|\boldsymbol{x}\|)$, is a closed set.

Definition 9.20 (**Minimal closed support *T*-invariant [S]**)**.** A *T*-invariant is a *minimal closed support T-invariant* if it is closed and has minimal support.

Definition 9.21 (**Liveness [B]**)**.** A transition is $t \in T$ is *live* if no matter what marking has been reached from $\boldsymbol{m}_0$ it is possible to ultimately fire transition t again. A Petri net is *live* under initial marking $\boldsymbol{m}_0$ if every transition is live under $\boldsymbol{m}_0$. An extensive discussion of liveness and related concepts is given in Murata [448].

Definition 9.22 (**Structural liveness [S]**)**.** A Petri net is *structurally live* if there exists an initial marking $\boldsymbol{m}_0$ for which the net is live.

Definition 9.23 (**Home state [B]**)**.** A marking $\boldsymbol{m}$ is a *home state* if for each marking in $\boldsymbol{m}' \in \mathcal{M}(\mathcal{PN}, \boldsymbol{m}_0)$, $\boldsymbol{m}$ is reachable from $\boldsymbol{m}'$, i.e., $\forall \boldsymbol{m}' \in \mathcal{M}(\mathcal{PN}, \boldsymbol{m}_0) : \boldsymbol{m} \in \mathcal{M}(\mathcal{PN}, \boldsymbol{m}')$.

Definition 9.24 (**Boundedness [B]**)**.** A Petri net is *k-bounded* or simply *bounded* if the number of tokens in each place does not exceed a finite number k for any marking in the reachability set $\mathcal{M}(\mathcal{PN}, \boldsymbol{m}_0)$.

Definition 9.25 (**Structural Boundedness [S]**)**.** A Petri net is *structurally bounded* if it is bounded for all initial markings.

Results

Result 9.26 (Murata [448])**.** A structurally bounded and structurally live Petri net is covered by both *P*-invariants and *T*-invariants.

Result 9.27 (Memmi and Roucairol [440])**.** There is a unique minimal *T*-invariant corresponding to a minimal support (*minimal support T-invariant*). Let $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^k$ denote the minimal support *T*-invariants. Any *T*-invariant $\boldsymbol{x}$ can be written as a linear combination of minimal support *T*-invariants:

$$\boldsymbol{x} = \sum_{i=1}^{k} \lambda_i \boldsymbol{x}^i$$

where $\lambda_i \in \mathbb{Q}^+$, $i = 1, \ldots, k$. The equivalent result holds for *P*-invariants.

Remark 9.28. Two remarks with respect to the decomposition result 9.27 of Memmi and Roucairol can be made. First, since the elements of minimal invariants are required to be non-negative, the minimal support invariants may be linearly dependent, so that there may exist more invariants than the dimension of the null space.

Second, for the decomposition to be in minimal support invariants it is essential that the weight factors λ_i are allowed to be rational numbers. If one restricts to integral weight factors, additional invariants may need to be added to the set of minimal support T-invariants to obtain a decomposition result. An extensive discussion on different decomposition results is provided by Krückeberg and Jaxy [364]. In this reference, efficient algorithms are also presented to obtain the sets of minimal T- and P-invariants from the incidence matrix $\boldsymbol{A}$.

Result 9.29 (Boucherie and Sereno [68])**.** A T-invariant $\boldsymbol{x}$ is a minimal closed support T-invariant if the firing sequence of $\boldsymbol{x}$ is *linear*, that is for each $t \in \|\boldsymbol{x}\|$ there is a unique $t' \in \|\boldsymbol{x}\|$ such that $\boldsymbol{O}(t) = \boldsymbol{I}(t')$. As a consequence $x_i \leq 1$, $i = 1, \ldots, M$. Conversely, if the firing sequence of a T-invariant $\boldsymbol{x}$ is linear, then $\boldsymbol{x}$ is a closed support T-invariant.

9.3.2 Stochastic Petri nets

Definition 9.30 (**Stochastic Petri net**)**.** A stochastic Petri net is a Petri net defined by the 5-tuple $\mathcal{SPN} = (P, T, I, O, Q)$, where (P, T, I, O) is a Petri net, and Q is a set of exponential firing rates $q(\boldsymbol{I}(t), \boldsymbol{O}(t); \boldsymbol{m} - \boldsymbol{I}(t))$ associated with the set of transitions $T = \{t_1, \ldots, t_M\}$ bringing marking $\boldsymbol{m}$ to $\boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{I}(t) + \boldsymbol{O}(t)$. Distributions associated with different transitions are independent. The firing execution policy of the stochastic Petri net is the race model.

Definition 9.31 (**Marked stochastic Petri net**)**.** A marked stochastic Petri net is a stochastic Petri net defined by the 6-tuple $(\mathcal{SPN}, \boldsymbol{m}_0) = (P, T, I, O, Q, \boldsymbol{m}_0)$, where $\boldsymbol{m}_0$ is the initial marking.

Definition 9.32 (**$\mathcal{S}\Pi$-net**)**.** A Π-net is a Petri net in which all transitions $t \in T$ are covered by minimal closed support T-invariants $\boldsymbol{x}^i, i = 1, \ldots, k$, that is for all $t \in T$ there exists an $i \in \{1, \ldots, k\}$ such that $t \in \|\boldsymbol{x}^i\|$ and $\|\boldsymbol{x}^i\|$ is a closed set. An $\mathcal{S}\Pi$-net is a stochastic Π-net.

There exist various firing execution policies for stochastic Petri nets. For an extensive discussion on these policies, see [418]. We assume that the firing execution policy follows a race model. As a consequence of the exponential firing times, the stochastic process describing the evolution of the Petri net is a time-homogeneous continuous-time Markov chain $\mathbf{X}$ at state space $\mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)$. Denote the transition rates of $\mathbf{X}$ by $Q_{\mathbf{X}} = (q(\boldsymbol{m}, \boldsymbol{m}'), \boldsymbol{m}, \boldsymbol{m}' \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0))$. To avoid anomalies, we assume the process is regular, that is, at most finitely many transitions can fire in finite time ([616], Chapter 2). It will be assumed that each transition of the Markov chain representing the Petri net is due to exactly one transition $t \in T$ that fires. Note that the firing of multiple transitions can be incorporated by adding extra transitions representing the combination of several transitions that fire with suitable firing rates.

The evolution of the Markov chain describing the stochastic Petri net is as follows. A transition t in marking $\boldsymbol{m}$ can be enabled only if $\boldsymbol{m} - \boldsymbol{I}(t) \in \mathbb{N}_0^N$. Furthermore, we will allow multiple transitions to have the same enabling condition, i.e.,

for $t_i \neq t_j$ it is allowed that $\boldsymbol{I}(t_i) = \boldsymbol{I}(t_j)$. Of course, the output bag will not be the same, otherwise these two transitions could be represented by only one. The rate

$$q(\boldsymbol{I}(t), \boldsymbol{O}(t); \boldsymbol{m} - \boldsymbol{I}(t)) \tag{9.1}$$

is associated with transition t bringing $\boldsymbol{m}$ to $\boldsymbol{m}' = \boldsymbol{m} - \boldsymbol{I}(t) + \boldsymbol{O}(t)$. Note that a transition from marking $\boldsymbol{m}$ to marking $\boldsymbol{m} - \boldsymbol{I}(t) + \boldsymbol{O}(t)$ may occur due to other transitions too. The total transition rate from marking $\boldsymbol{m}$ to marking $\boldsymbol{m}'$ is therefore

$$q(\boldsymbol{m}, \boldsymbol{m}') = \sum_{\{\boldsymbol{n} \in \mathbb{N}_0^N,\ t \in T:\ \boldsymbol{n} + \boldsymbol{I}(t) = \boldsymbol{m},\ \boldsymbol{n} + \boldsymbol{O}(t) = \boldsymbol{m}'\}} q(\boldsymbol{I}(t), \boldsymbol{O}(t); \boldsymbol{n}). \tag{9.2}$$

When analyzing the Markov chain $\mathbf{X}$ describing the behavior of a stochastic Petri net, it will be convenient to aggregate transitions with identical input bag to one transition with a probabilistic output bag. In that case, all transitions, say $t_{i_1}, \ldots, t_{i_k}$ with identical input bag are aggregated into a single transition t. The output bag of this new transition is probabilistic, with the probability that output bag $\boldsymbol{O}(t_{i_j})$ occurs determined by the original firing rates, so that:

$$q(\boldsymbol{I}(t), \boldsymbol{O}(t); \boldsymbol{m} - \boldsymbol{I}(t)) = \mu(t; \boldsymbol{m} - \boldsymbol{I}(t)) p(\boldsymbol{I}(t), \boldsymbol{O}(t); \boldsymbol{m} - \boldsymbol{I}(t)). \tag{9.3}$$

where $\mu(t; \boldsymbol{m} - \boldsymbol{I}(t)) = \sum_{j=1}^{k} q(\boldsymbol{I}(t_{i_j}), \boldsymbol{O}(t_{i_j}); \boldsymbol{m} - \boldsymbol{I}(t_{i_j}))$ is the total firing rate and $p(\boldsymbol{I}(t), \boldsymbol{O}(t_{i_j}); \boldsymbol{m} - \boldsymbol{I}(t)) = q(\boldsymbol{I}(t_{i_j}), \boldsymbol{O}(t_{i_j}); \boldsymbol{m} - \boldsymbol{I}(t_{i_j}))/\mu(t; \boldsymbol{m} - \boldsymbol{I}(t))$ is the probability of selecting a specific output bag $\boldsymbol{O}(t_{i_j})$.

We are interested in calculating the steady-state behavior of the continuous-time Markov chain $\mathbf{X}$ modeling the marked stochastic Petri net $(\mathcal{SPN}, \boldsymbol{m}_0)$. From standard Markov theory we know that $\mathbf{X}$ is irreducible and positive recurrent if and only if a unique collection of positive numbers $\pi = (\pi(\boldsymbol{m}), \boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0))$ summing to unity, exists satisfying the *global balance equations*,

$$\sum_{\boldsymbol{m}' \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)} \{\pi(\boldsymbol{m}) q(\boldsymbol{m}, \boldsymbol{m}') - \pi(\boldsymbol{m}') q(\boldsymbol{m}', \boldsymbol{m})\} = 0 \ , \boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0). \tag{9.4}$$

This $\pi = (\pi(\boldsymbol{m}), \boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0))$ is called the *equilibrium distribution*.

As the Markov chain is chosen such that it describes the evolution of the stochastic Petri net under consideration, irreducibility and positive recurrence properties necessary to obtain a unique equilibrium distribution for the Markov chain should preferably be characterized directly from the Petri net structure.

The state space of a Markov chain $\mathbf{X}$ partitions in communicating classes [509]. Because we are interested in the steady state behavior of $\mathbf{X}$ we can analyze the process at each class separately. Moreover, we are not interested in transient classes, as transient states will vanish in the equilibrium distribution of the stochastic Petri net. Thus, we will focus on stochastic Petri nets of which the corresponding Markov chain $\mathbf{X}$ is irreducible.

To prevent the presence of transient classes, we restrict ourselves to bounded Petri nets that are live and therefore covered by T-invariants. If the Petri net is

live and has a home state, then $\mathbf{X}$ is irreducible. (Note that irreducibility of the Markov chain is called reversibility in the Petri net literature [448]. The notion of reversibility for Petri nets should not be confused with the notion of reversibility for Markov chains [347]).

If the reachability set is finite, positive recurrence follows from irreducibility. Otherwise, for $\mathbf{X}$ to be stable additional assumptions on the transition rates are required to ensure that the rate at which tokens are created is smaller then the rate at which they are destroyed. This problem is for example addressed in [217]. To avoid non-regularity, we restrict our attention to stochastic Petri nets with a finite reachability set, thus to structurally bounded nets. By Result 9.26, for a live net to be structurally bounded, the net must be covered by P-invariants.

A live Petri net is structurally live. A complete characterization of structural liveness for a general Petri net is unknown [448]. Liveness and boundedness are not related to the existence of a home state [448] for general net structures. It is beyond the scope of this dissertation to provide a complete overview for general Petri nets (see [199] and [448] for elaborate discussions). For $\mathcal{S}\Pi$-nets (see Definition 9.32), in Theorem 10.13 we will provide a complete characterization of structurally liveness and existence of a home state. Note that also in this case, for a specific initial marking liveness still needs to be checked, which may be a cumbersome problem (see Haddad et al. [270] for some exploratory results).

9.4 Literature

Product form results exist on different levels. In the classical product form result the equilibrium distribution of a network can be expressed as a product over the nodes of the network. In this section we provide a survey of such results for queueing networks, stochastic process algebras and stochastic Petri nets in Sections 9.4.1, 9.4.2, and 9.4.3, respectively. A more general product form result is when the equilibrium distribution of a network is a (normalized) product over the marginal distribution of subnets. A survey of such decomposition results will be provided in Section 9.4.4.

9.4.1 Product form results for queueing networks

For queueing networks an important analytical result is the product form equilibrium distribution for the number of customers at the stations. The basis of the development of product form literature is given by Jackson [330]. Jackson's product form states that the equilibrium distribution of the queueing network is the product of the marginal distributions at the stations of the queueing network. Product form results for closed queueing networks, networks in which a fixed number of customers is present, were obtained by Gordon and Newell [243]. The results of Jackson [330] and Gordon and Newell [243] were proven on the basis of global balance.

The concept of partial balance as the basis of product form was introduced in [630, 631]. These results were generalized to Kelly-Whittle networks (see, e.g., [347, 632]), networks with job-types and various service disciplines (see, e.g., [31,

310, 582]) and to batch routing (see, e.g., [69, 297, 299]) and discrete-time networks (see, e.g., [143]). A different approach for obtaining product form equilibrium distributions is based on the notion of quasi-reversibility (see, e.g., [115, 347, 446]).

9.4.2 Product form results for stochastic Petri nets

For stochastic Petri nets, the first product form results for the number of tokens at the places were obtained by Lazar and Robertazzi [384] for the class of stochastic Petri nets consisting of 'linear task sequences', a number of tasks that must be executed consecutively. Since these first results, considerable extensions have been derived by several authors. In a series of papers, Henderson et al. [296, 298, 300] translated and extended product form results for batch routing queueing networks to stochastic Petri nets, which are equivalent to batch routing queueing networks at the level of the underlying stochastic process.

The starting point for the analysis of product form stochastic Petri nets is the assumption that a solution exists for the 'routing chain', a set of linear equations similar to the traffic equations for queueing networks. The product form results for stochastic Petri nets obtained in [296, 298, 300] were based on the assumption that a positive solution exists for the routing chain. Necessary conditions for such a solution to exist were provided in Henderson et al. [296].

A full characterization of the structure of stochastic Petri nets necessary and sufficient for the existence of a positive solution for the routing chain was obtained in [66, 182]: all transitions of the Petri net should be covered by 'closed support T-invariants'. This new type of T-invariant was also introduced in [66, 182] and is a T-invariant that closely resembles the 'task sequences' used by Lazar and Robertazzi [384]. As such, the existence of a solution for the routing chain was completely characterized on the basis of the structure of the Petri net. This class of stochastic Petri nets was later denoted as $\mathcal{S}\Pi$-nets by Haddad et al [270].

For an $\mathcal{S}\Pi$-net, Coleman et al. [130] were the first to formulate an additional requirement sufficient for product form in stochastic Petri net by a numerical condition on the transition rates. Haddad et al. [270] and Mairesse and Nguyen [406] established characterizations of $\mathcal{S}\Pi$-nets with a product form solution irrespective of the values of the transition rates. Haddad et al. achieved this via the concept of $\mathcal{S}\Pi^2$-nets and Mairesse and Nguyen via the concept of 'zero-deficiency' $\mathcal{S}\Pi$-nets. The conditions of Coleman et al., Haddad et al. and Mairesse and Nguyen are algebraic conditions which lack intuition in terms of Petri net structure. *In Chapter 10, we unify these results by the concept of group-local-balance and extends these results by formulating all product form results in terms of T-invariants.*

9.4.3 Product form for stochastic process algebras

The stochastic process algebras formalism is build upon the classical process algebras during the 1990s to include actions requiring a random time. The principle of process algebras is that complex systems are defined by a composed collection of

agents who execute actions, which may or may not be concurrent. Various different languages of stochastic process algebras were introduced. Although most product form results are formulated in the paradigm of *Performance Evaluation Process Algebra* (PEPA), defined by Hillston in [305], the results can easily be generalized to any of the other stochastic process algebras.

A comprehensive survey of product form results for stochastic process algebras can be found in the PhD thesis Marin [414]. Marin distinguishes between various types of product form results: models based on reversibility (e.g., [306]), models based on quasi-reversibility (e.g., [290]), models based on the product form results for stochastic Petri nets by Henderson et. al [296] and Coleman et al. [130] (e.g., [523]) and models based on the *Reversed Compound Agent Theorem* (RCAT) theorem and its extensions (e.g., [287, 288, 289]). In addition, models based on the cooperating Markov chains of the form presented by Boucherie in [64] are distinguished (e.g., [289, 307]).

9.4.4 Decomposition

A network can be decomposed if its stationary distribution factorizes into the stationary distributions of the nodes of which the network is comprised; the network is then of product form. Apart from the theoretical interest, decomposition results are also of substantial practical importance: finding the stationary distribution of an entire network usually requires an enormous computational effort, whereas the stationary distribution of a single node can be found relatively easily. The first, and perhaps most famous, decomposition results for queueing networks have been reported by Jackson [330]: the classical Jackson product form result. Decomposition of networks into subnetworks have been a topic of research for queueing networks. Two streams of literature have been developed in parallel: results based on partial balance (e.g., [70, 79, 114, 311, 361]) and results based on quasi-reversibility (e.g., [65, 78, 615, 617]). Recently, in a setting of general stochastic processes, these results have been unified and extended in [115, 318].

For stochastic Petri nets decomposition results were initialized by Lazar and Robertazzi [385] for connected subnets of task sequences and were extended by Boucherie [64] in the framework of competing Markov chains. Frosch and Natarajan [222, 223] derived product form results for so-called closed synchronized systems of stochastic sequential processes, a class of Petri nets in which state machines are synchronized via buffer places. The results in these references may also be interpreted as composition results since the networks are essentially obtained by composing subnets in to a larger net, similar to the composition structure of stochastic process algebras. As such, no procedure is provided in the literature to algorithmically characterize subnets in a given stochastic Petri net and to verify whether product form holds. *In Chapters 11 and 12, we present decomposition results for stochastic Petri nets completely formulated on their structure in terms of P- and T-invariants.*

CHAPTER 10

Structural Characterization of Product Form

10.1 Introduction

In this chapter, we survey the various structural results that are known for stochastic Petri nets with a product form equilibrium distribution over the number of tokens at the places and rephrases all these results in terms of T-invariants. In addition, we unify and extend the product form results for stochastic Petri nets by showing that *group-local-balance* can be identified as the concept underlying all these structural results and we provide additional structural implications and an intuitive explanation of the known and new results, all based on T-invariants only.

The chapter is organized as follows. Section 10.2 translates product form results for batch routing queueing networks based on the group-local-balance concept into Petri net terminology. These results, presented on the Markov chain level, provide the basis for Section 10.3, in which structural Petri net implications are discussed. Section 10.3 concludes with an algorithm to verify whether a specific stochastic Petri net has a product form equilibrium distribution, and if so, to construct this product form. To provide an illustration of the results, in Section 12.4 several examples of product form stochastic Petri nets are presented.

10.2 Group-local-balance

In this section, we analyze the Markov chain $\mathbf{X}$ of an $\mathcal{SPN}$. Boucherie and Van Dijk [69] presented the group-local-balance concept as the basis for the analysis of product form batch routing queueing networks. Here, we translate the definitions and results of [69] into Petri net terminology, and we show that group-local-balance allows us to calculate the steady state distribution of an $\mathcal{SPN}$. This will be the foundation to investigate the structural Petri net implications of group-local-balance in Section 10.3.

Inserting the transition rates (9.2) into the global balance equations (9.4) yields that a distribution π at $\mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)$ is the unique equilibrium distribution if for all $\boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)$:

$$\sum_{\{\boldsymbol{n},\, t,\, t' \in T : \boldsymbol{n}+\boldsymbol{I}(t)=\boldsymbol{n}+\boldsymbol{O}(t')=\boldsymbol{m}\}} \left\{ \pi(\boldsymbol{m}) q(\boldsymbol{I}(t), \boldsymbol{O}(t); \boldsymbol{n}) - \pi(\boldsymbol{n}+\boldsymbol{I}(t')) q(\boldsymbol{I}(t'), \boldsymbol{O}(t'); \boldsymbol{n}) \right\} = 0.$$

A distribution satisfying these equations for fixed combinations of residual marking $\boldsymbol{n}$ and input bag $\boldsymbol{I}(t)$ is the unique equilibrium distribution. This form of *local balance* is introduced in [69] as *group-local-balance*.

Definition 10.1 (Group-local-balance). A measure ϕ satisfies *group-local-balance* (GLB) if, for all fixed residual markings $\boldsymbol{n}$ and for all fixed input bags $\boldsymbol{I}(t)$, such that $\boldsymbol{n}+\boldsymbol{I}(t) \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)$:

$$\sum_{\{t' \in T : \boldsymbol{I}(t')=\boldsymbol{I}(t)\}} \phi(\boldsymbol{n}+\boldsymbol{I}(t'))q(\boldsymbol{I}(t'),\boldsymbol{O}(t');\boldsymbol{n}) = \sum_{\{t' \in T : \boldsymbol{O}(t')=\boldsymbol{I}(t)\}} \phi(\boldsymbol{n}+\boldsymbol{I}(t'))q(\boldsymbol{I}(t'),\boldsymbol{O}(t');\boldsymbol{n}). \quad (10.1)$$

Summation of the group-local-balance equations over all $\boldsymbol{n}, \boldsymbol{I}(t)$ such that $\boldsymbol{n}+\boldsymbol{I}(t) = \boldsymbol{m}$ gives the global balance equations. The Markov chain $\mathbf{X}$ has the GLB-property if the equilibrium distribution π satisfies (10.1).

GLB expresses that under a given residual marking the rate at which input bag $\boldsymbol{I}(t)$ is absorbed is balanced by the rate at which exactly $\boldsymbol{I}(t)$ is formed. Obviously, the group-local-balance equations are generally more restrictive than the global balance equations. GLB requires that $\boldsymbol{I}(t)$ is an output bag of a transition t'. Also, GLB requires the output bag of a transition t to be the input bag for another transition t'.

Lemma 10.2. If the Markov chain $\mathbf{X}$ of an $\mathcal{SPN}$ satisfies GLB, then $\mathcal{R}(T)$ is a closed set.

Proof. From the group-local-balance equations (10.1) it is seen that if $\boldsymbol{I}(t)$ is an input bag of a transition that is enabled in an arbitrary marking $\boldsymbol{m}$, then, if GLB holds, $\boldsymbol{I}(t)$ must also be an output bag of a transition t'. If there is no such transition t', the left hand side of (10.1) would be positive while the right hand side is zero, which contradicts GLB.

Similarly, if $\boldsymbol{O}(t')$ is an output bag of a transition that is enabled in an arbitrary marking $\boldsymbol{m}$, then, if GLB holds, $\boldsymbol{O}(t')$ must also be an input bag of a transition t. If there is no such transition t, the right hand side of (10.1) would be positive while the left hand side is zero, which contradicts GLB. □

Following [69], let us introduce the concepts of the *local state space* and the *local irreducible sets*. For a fixed $\boldsymbol{n}$ the local state space $V(\boldsymbol{n})$ is the state space of the Markov chain with transition rates $q(\boldsymbol{I}(t),\boldsymbol{O}(t);\boldsymbol{n})$ restricted to $\mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)$. So $V(\boldsymbol{n})$ consists of all states $\boldsymbol{n}+\boldsymbol{I}(t)$ and $\boldsymbol{n}+\boldsymbol{O}(t)$, for which $q(\boldsymbol{I}(t),\boldsymbol{O}(t);\boldsymbol{n}) > 0$. Let $V_i(\boldsymbol{n})$ denote the local irreducible sets in $V(\boldsymbol{n})$ with respect to the Markov chain with transition rates $q(\boldsymbol{I}(t),\boldsymbol{O}(t);\boldsymbol{n})$ for fixed $\boldsymbol{n}$. A state $\boldsymbol{m}$ may be element of different local state spaces $V(\boldsymbol{n})$, so that transitions from one local state space to another are possible. It is not uncommon that $V(\boldsymbol{n})$ consists of multiple local irreducible sets $V_i(\boldsymbol{n}), i \in \{1,\ldots,k(\boldsymbol{n})\}$, which is shown in [69] via an example. In addition, it is shown that if a Markov chain satisfies GLB, the local state spaces $V(\boldsymbol{n})$ consist only

of irreducible sets, which guarantees:

$$V(\boldsymbol{n}) = \bigcup_{i=1}^{k(\boldsymbol{n})} V_i(\boldsymbol{n}).$$

Now, it follows that, if the Markov chain $\mathbf{X}$ of an $\mathcal{SPN}$ net has the GLB property, then for any fixed $\boldsymbol{n}$ for which $V(\boldsymbol{n}) \neq \emptyset$ and $i \in \{1,\ldots,k(\boldsymbol{n})\}$ the following set of equations has a unique positive solution up to a multiplicative constant; for $\boldsymbol{n}+\boldsymbol{I}(t) \in V_i(\boldsymbol{n})$:

$$x(\boldsymbol{I}(t);\boldsymbol{n}) \sum_{t' \in T} q(\boldsymbol{I}(t),\boldsymbol{I}(t');\boldsymbol{n}) = \sum_{t' \in T} x(\boldsymbol{I}(t');\boldsymbol{n}) q(\boldsymbol{I}(t'),\boldsymbol{I}(t);\boldsymbol{n}). \tag{10.2}$$

These local solutions per communicating class can be used to characterize the equilibrium distribution π, by translating these solutions to the global state space. To this end, an additional process with transition rate $\bar{q}$ is defined. For any Markov chain $\mathbf{X}$ at $\mathcal{M}(\mathcal{SPN},\boldsymbol{m}_0)$ that satisfies the equations (10.2) the $\bar{q}$-process can be defined. However, such a Markov chain does not necessarily satisfy the GLB property. To point out in when this relation does hold, [69] introduces the concept of strong reversibility.

Definition 10.3 (**$\bar{q}$-process**)**.** If for any fixed $\boldsymbol{n}$ for which $V(\boldsymbol{n}) \neq \emptyset$, the system (10.2) has for $i \in \{1,\ldots,k(\boldsymbol{n})\}$ a unique positive solution $\{x(\boldsymbol{I}(t);\boldsymbol{n}) \mid \boldsymbol{n}+\boldsymbol{I}(t) \in V_i(\boldsymbol{n})\}$ up to a multiplicative constant, then the following process, called the $\bar{q}$-*process*, can be defined.

For any $\boldsymbol{n}, i \in \{1,\ldots,k(\boldsymbol{n})\}$, and $\boldsymbol{n}+\boldsymbol{I}(t), \boldsymbol{n}+\boldsymbol{I}(t') \in V_i(\boldsymbol{n})$, for which $q(\boldsymbol{I}(t),\boldsymbol{I}(t');\boldsymbol{n}) > 0$ or $q(\boldsymbol{I}(t'),\boldsymbol{I}(t);\boldsymbol{n}) > 0$

$$\frac{\bar{q}(\boldsymbol{I}(t),\boldsymbol{I}(t');\boldsymbol{n})}{\bar{q}(\boldsymbol{I}(t'),\boldsymbol{I}(t);\boldsymbol{n})} = \frac{x(\boldsymbol{I}(t'),\boldsymbol{n})}{x(\boldsymbol{I}(t),\boldsymbol{n})}, \tag{10.3}$$

and otherwise

$$\bar{q}(\boldsymbol{I}(t),\boldsymbol{I}(t');\boldsymbol{n}) = 0.$$

Definition 10.4 (**Strong reversibility**)**.** The $\bar{q}$-process is called *strongly reversible* at $\mathcal{M}(\mathcal{SPN},\boldsymbol{m}_0)$ if for all $\boldsymbol{n}$ for which $V(\boldsymbol{n}) \neq \emptyset$ and $i \in \{1,\ldots,k(\boldsymbol{n})\}$, the equilibrium distribution $\bar{\pi}$ satisfies for $\boldsymbol{n}+\boldsymbol{I}(t), \boldsymbol{n}+\boldsymbol{I}(t') \in V_i(\boldsymbol{n})$:

$$\bar{\pi}(\boldsymbol{n}+\boldsymbol{I}(t))\bar{q}(\boldsymbol{I}(t),\boldsymbol{I}(t');\boldsymbol{n}) = \bar{\pi}(\boldsymbol{n}+\boldsymbol{I}(t'))\bar{q}(\boldsymbol{I}(t'),\boldsymbol{I}(t);\boldsymbol{n}).$$

Theorem 10.5 ([69])**.** The equilibrium distribution of a Markov chain $\mathbf{X}$ at satisfies GLB if and only if the $\bar{q}$-process is defined and is strongly reversible at $\mathcal{M}(\mathcal{SPN},\boldsymbol{m}_0)$. Moreover, with $\bar{\pi}$ its equilibrium distribution, for all $\boldsymbol{m} \in \mathcal{M}(\mathcal{SPN},\boldsymbol{m}_0)$: $\pi(\boldsymbol{m}) = \bar{\pi}(\boldsymbol{m})$. Finally, π satisfies GLB if and only if for an arbitrary reference state $\boldsymbol{m}_0$, and all $\boldsymbol{m} \in \mathcal{M}(\mathcal{SPN},\boldsymbol{m}_0)$

$$\pi(\boldsymbol{m}) = \pi(\boldsymbol{m}_0) \prod_{k=0}^{s} \frac{\bar{q}(\boldsymbol{I}(t_k),\boldsymbol{I}(t'_k);\boldsymbol{n}_k)}{\bar{q}(\boldsymbol{I}(t'_k),\boldsymbol{I}(t_k);\boldsymbol{n}_k)}, \tag{10.4}$$

for all firing sequences of the form (such that the denominator of (10.4) is positive)

$$\begin{aligned}\boldsymbol{m}_0 = \boldsymbol{n}_0 + I(t_0) \to \boldsymbol{n}_0 + I(t_0') = \boldsymbol{n}_1 + \boldsymbol{I}(t_1) \to \boldsymbol{n}_1 + I(t_1') = \ldots \to \\ \ldots = \boldsymbol{n}_s + \boldsymbol{I}(t_s) \to \boldsymbol{n}_s + \boldsymbol{I}(t_s') = \boldsymbol{n}_{s+1} + \boldsymbol{I}(t_{s+1}) = \boldsymbol{m}.\end{aligned}$$

Corollary 10.6. The equilibrium distribution π satisfies GLB if and only if for $\boldsymbol{n}, \boldsymbol{I}(t)$ and $\boldsymbol{I}(t')$ such that $\boldsymbol{n}+\boldsymbol{I}(t), \boldsymbol{n}+\boldsymbol{I}(t') \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)$, for which $q(\boldsymbol{I}(t), \boldsymbol{I}(t'); \boldsymbol{n}) > 0$

$$\frac{\pi(\boldsymbol{n}+\boldsymbol{I}(t))}{\pi(\boldsymbol{n}+\boldsymbol{I}(t'))} = \frac{x(\boldsymbol{I}(t); \boldsymbol{n})}{x(\boldsymbol{I}(t'); \boldsymbol{n})}. \tag{10.5}$$

Corollary 10.6 provides the relation between the equilibrium distribution π and the local solutions $x(\boldsymbol{n}; \boldsymbol{I}(t))$. Note that (10.5) is a condition for $\boldsymbol{n}$, $\boldsymbol{I}(t)$ and $\boldsymbol{I}(t')$ such that $\boldsymbol{n}+\boldsymbol{I}(t)$ and $\boldsymbol{n}+\boldsymbol{I}(t')$ are within a single local irreducible set $V_i(\boldsymbol{n})$, and it relates the ratio $x(\boldsymbol{I}(t); \boldsymbol{n})/x(\boldsymbol{I}(t'); \boldsymbol{n})$ to the ratio $\pi(\boldsymbol{n}+\boldsymbol{I}(t))/\pi(\boldsymbol{n}+\boldsymbol{I}(t'))$. For a firing sequence from marking $\boldsymbol{m}$ to $\boldsymbol{m}'$ that traverses multiple local irreducible sets $V_j(\boldsymbol{n}_j)$, $j = 1, \ldots, s$, for each transition in this firing sequence (10.5) is imposed. The latter implies that if there exist multiple firing sequences from $\boldsymbol{m}$ to $\boldsymbol{m}'$ additional restrictions on the ratios $\bar{q}(\boldsymbol{I}(t_k), \boldsymbol{I}(t_k'); \boldsymbol{n}_k)/\bar{q}(\boldsymbol{I}(t_k'), \boldsymbol{I}(t_k); \boldsymbol{n}_k)$ in (10.4) are implied to obtain consistency in the ratio $\pi(\boldsymbol{m})/\pi(\boldsymbol{m}')$ in (10.4). In Section 10.3, the impact of these conditions at the Petri net level will be studied in detail.

This section has described results on the Markov chain level. Reversibility of the $\bar{q}$-process provides a way to 'build' the solution $\bar{\pi}(\boldsymbol{m})$, following any path to $\boldsymbol{m}$ from the initial marking $\boldsymbol{m}_0$. To understand and exploit the results on the Petri net level, in the next section, we will investigate the translation of these characteristics to the stochastic Petri nets and in particular present the implications for the stochastic Petri net structure. The key ingredients of that analysis will be the local irreducible sets and ratio condition of Corollary 10.6.

10.3 Product form

In this section, we will show that stochastic Petri nets with marking-independent firing rates for which group-local-balance holds have a steady state distribution that is a product over the places of the network. Therefore, we are interested in the necessary and sufficient structural properties of Petri nets that are required to obtain group-local-balance.

The first structural condition was already presented in Lemma 10.2: the set of input and output bags $\mathcal{R}(T)$ is a closed set. In Section 10.3.1, this condition is extended to 'each transition has to be covered by a minimal closed support T-invariant', i.e., the $\mathcal{SPN}$ has to be an $\mathcal{S}\Pi$-net. To this end, it is shown that the local irreducible sets defined in Section 10.2 are sets of minimal closed support T-invariants. Section 10.3.2 shows that an $\mathcal{S}\Pi$-net does not necessarily have a product form solution. The additional relation between states can be found by tracing

closed support T-invariants. This observation forms the key to formulate the additional requirements to obtain a characterization of product form stochastic Petri nets. Section 10.3.3 identifies the structural characteristics of $\mathcal{S}\Pi$-nets for which a product form equilibrium distribution can be concluded without considering the numerical values of the transition rates and nets for which these values have to satisfy specific conditions. This subsection concludes with an algorithm to verify whether a specific $\mathcal{SPN}$ has a product form equilibrium distribution, and if so, to construct this product form. Section 12.4 provides several insightful examples of product form $\mathcal{SPN}$s.

The Markov chain $\mathbf{X}$ on state space $\mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)$ modeling the Petri net with marking-independent firing rates has transition rates

$$q(\boldsymbol{I}(t), \boldsymbol{O}(t); \boldsymbol{m} - \boldsymbol{I}(t)) = \mu(t) p(\boldsymbol{I}(t), \boldsymbol{O}(t)) \mathbb{1}_{(m(n) \geq I_n(t),\, n=1,\ldots,N)}. \tag{10.6}$$

Observe that for the nets with transition rates (10.6) the condition $m(n) \geq I_n(t)$, $n = 1, \ldots, N$, is necessary and sufficient for transition t to be enabled in marking $\boldsymbol{m}$.

10.3.1 Routing chain and minimal closed support T-invariants

Under marking independent transition rates the equations (10.2) are equivalent for all $\boldsymbol{n} + \boldsymbol{I}(t) \in V_i(\boldsymbol{n})$, which can be seen from inserting (10.6) in (10.2), for all $\boldsymbol{n} + \boldsymbol{I}(t) \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)$:

$$\begin{aligned} x(\boldsymbol{I}(t); \boldsymbol{n}) \sum_{t' \in T} \mu(t) p(\boldsymbol{I}(t), \boldsymbol{I}(t')) \mathbb{1}_{(m(n) \geq I_n(t),\, n=1,\ldots,N)} \\ = \sum_{t' \in T} x(\boldsymbol{I}(t'); \boldsymbol{n}) \mu(t') p(\boldsymbol{I}(t'), \boldsymbol{I}(t)) \mathbb{1}_{(m(n) \geq I_n(t'),\, n=1,\ldots,N)}. \end{aligned} \tag{10.7}$$

Considering (10.7) for all residual markings $\boldsymbol{n}$ and input bags $\boldsymbol{I}(t)$ and local irreducible sets $V_i(\boldsymbol{n})$ such that $\boldsymbol{n} + \boldsymbol{I}(t) \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)$, exposes that the set of equations of the form (10.7) only differ in the local irreducible sets $V_i(\boldsymbol{n})$ ($i \in 1, \ldots, k(\boldsymbol{n})$) being enabled or disabled. Therefore, if the equilibrium distribution π satisfies GLB, then for each $\boldsymbol{n} + \boldsymbol{I}(t) \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)$ equation (10.7) has a unique positive solution $x(\boldsymbol{I}(t); \boldsymbol{n}) := y(\boldsymbol{I}(t))$.

This implies that a positive solution can be found to the global balance equations of a Markov chain which is defined by Henderson et al. as the *routing chain* [296]. Define the Markov chain $\mathbf{Y} = (Y(t), t \geq 0)$ on finite state space $S = \{\boldsymbol{I}(t), t \in T\}$ with transition rates $q_{\mathbf{Y}}(\boldsymbol{I}(t), \boldsymbol{I}(t')) = \mu(t) p(\boldsymbol{I}(t), \boldsymbol{I}(t'))$. The global balance equations for $\mathbf{Y}$ are, for $t \in T$,

$$\sum_{t' \in T} \{ y(\boldsymbol{I}(t)) \mu(t) p(\boldsymbol{I}(t), \boldsymbol{I}(t')) - y(\boldsymbol{I}(t')) \mu(t') p(\boldsymbol{I}(t'), \boldsymbol{I}(t)) \} = 0. \tag{10.8}$$

These global balance equations for Markov chain $\mathbf{Y}$ are state independent versions of the group-local-balance equations (10.2). The definition of the routing chain relies on the condition that $\mathcal{R}(T)$ is a closed set, so that for all $t \in T$, $\boldsymbol{I}(t) = \boldsymbol{O}(t')$ for some t' and therefore $p(\boldsymbol{I}(t), \boldsymbol{I}(t')) = p(\boldsymbol{I}(t), \boldsymbol{O}(t))$ is well-defined.

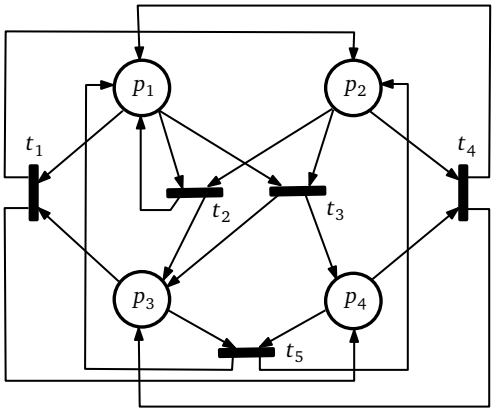


Figure 10.1: Petri net for which $\mathcal{R}(T)$ is a closed set.

Observe that GLB cannot hold if no positive solution for the routing chain can be found. Therefore, in the following, we first investigate the structural conditions under which a positive solution for the routing chain exists. The condition that $\mathcal{R}(T)$ is a closed set is necessary for a solution $\mathbf{Y}$ to exist. This condition is exactly the condition that Henderson et al. impose in Corollary 1 of [296] on the $\mathcal{SPN}$s they consider. In their further analysis, they assume a positive solution for the routing chain exists; an assumption which is usually made in the literature. The following example, taken from [66], shows that the closedness of $\mathcal{R}(T)$ is not a sufficient condition for GLB to hold.

Example 10.7. Consider the $\mathcal{SPN}$ depicted in Figure 10.1. $\boldsymbol{I}(t_1) = (1,0,1,0)$, $\boldsymbol{I}(t_2) = (1,1,0,0)$, $\boldsymbol{I}(t_3) = (1,1,0,0)$, $\boldsymbol{I}(t_4) = (0,1,0,1)$, $\boldsymbol{I}(t_5) = (0,0,1,1)$ and $\boldsymbol{O}(t_1) = (0,1,0,1)$, $\boldsymbol{O}(t_2) = (1,0,1,0)$, $\boldsymbol{O}(t_3) = (0,0,1,1)$, $\boldsymbol{O}(t_4) = (1,0,1,0)$, $\boldsymbol{O}(t_5) = (1,1,0,0)$, which shows that $\mathcal{R}(T)$ is a closed set. Since $\boldsymbol{I}(t_2) = \boldsymbol{I}(t_3)$, the state space of the routing chain is $S = \{\boldsymbol{I}(t_1), \boldsymbol{I}(t_2), \boldsymbol{I}(t_4), \boldsymbol{I}(t_5)\}$, and the solution for the routing chain (10.8) is (up to a multiplicative constant)

$$y(\boldsymbol{I}(t_1)) = 1/\mu_1, y(\boldsymbol{I}(t_4)) = 1/\mu_4, y(\boldsymbol{I}(t_2)) = y(\boldsymbol{I}(t_3)) = y(\boldsymbol{I}(t_5)) = 0,$$

which shows that closedness of $\mathcal{R}(T)$ is not sufficient for a *positive* solution for the routing chain. □

In Example 10.7, $\mathbf{Y}$ does not partition in irreducible classes, because $S_1 = \{\boldsymbol{I}(t_2), \boldsymbol{I}(t_5)\}$ is a transient class. Boucherie and Sereno [68] present a necessary and sufficient condition: for an $\mathcal{SPN}$ a positive solution for the routing chain exists if and only if all transitions $t \in T$ are covered by minimal closed support T-invariants, i.e., it is an $\mathcal{S}\Pi$-net. They prove this by showing that only in this case does the state space of the Markov chain $\mathbf{Y}$ partition into irreducible sets.

Obviously, the condition of the $\mathcal{SPN}$ to be an $\mathcal{S}\Pi$-net implies that $\mathcal{R}(T)$ is a closed set. In addition to the closedness condition, in an $\mathcal{S}\Pi$-net transitions t, s

with $\boldsymbol{O}(t) = \boldsymbol{I}(s)$ are elements of the support of a single minimal closed support T-invariant. Returning to example 10.7 illustrates this essential extension.

Example 10.7 revisited. From the incidence matrix

$$\boldsymbol{A} = \begin{pmatrix} -1 & 0 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 \\ -1 & 1 & 1 & 1 & -1 \\ 1 & 0 & 1 & -1 & -1 \end{pmatrix}$$

we obtain that this net has 3 minimal support T-invariants: $\boldsymbol{x}^1 = (10010)$, $\boldsymbol{x}^2 = (00101)$, $\boldsymbol{x}^3 = (12001)$, of which $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$ have closed support, but $\boldsymbol{x}^3$ does not have closed support. Since transition t_2 is contained in $\|\boldsymbol{x}^3\|$ only, t_2 is not covered by a minimal closed support T-invariant, which contradicts the definition of an $\mathcal{S}\Pi$-net. This explains why no positive solution for the routing chain exists. □

Observe that the essential characteristic of an $\mathcal{S}\Pi$-net is that all transitions are contained in a *closed* support T-invariant. The condition that all transitions are covered by minimal support T-invariants (closed or not closed) is a natural assumption if one is interested in the equilibrium or stationary distribution of a stochastic Petri net (see Section 9.3.2). A Petri net consisting of minimal closed support T-invariants is the natural extension of a state machine.

To obtain the partitioning of $\mathbf{Y}$ into irreducible classes, Boucherie and Sereno [68] provide a decomposition of the transitions of the Petri net into equivalence classes based on the characterization of minimal closed support T-invariants that are connected by having an input bag in common. By this equivalence class decomposition, the global balance equations of the routing chain (10.8) decompose into disjoint sets of equations, one set of equations for each equivalence class of connected T-invariants. The equivalence relation is defined by analogy with a similar equivalence relation introduced in Frosch and Natarajan [223] for cyclic state machines.

Assume that the minimal support T-invariants $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^h$ are numbered such that $C\ell T := \{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^k\}$ is the set of minimal closed support T-invariants ($k \leq h$).

Definition 10.8 (**Common input bag relation** [68])**.** Let $\boldsymbol{x}, \boldsymbol{x}' \in C\ell T$. The T-invariants $\boldsymbol{x}, \boldsymbol{x}'$ are in common input bag relation (notation: $\boldsymbol{x}\ CI\ \boldsymbol{x}'$) if there exist $t \in \|\boldsymbol{x}\|$, $t' \in \|\boldsymbol{x}'\|$ such that $\boldsymbol{I}(t) = \boldsymbol{I}(t')$. The relation CI^* is the transitive closure[1] of CI.

Definition 10.9 (**Common input bag class** [68])**.** The common input bag class $CI(\boldsymbol{x})$ is the equivalence class of $\boldsymbol{x} \in C\ell T$, that is $CI(\boldsymbol{x}) = \{\boldsymbol{x}' | \boldsymbol{x}\ CI^*\ \boldsymbol{x}'\}$.

The common input bag relation characterizes the irreducible sets of the routing chain. Before we specify this, let us first introduce some additional notation.

[1]The transitive closure of a relation is defined as follows: if $\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{x}'' \in C\ell T$, and $\boldsymbol{x}\ CI\ \boldsymbol{x}'$, $\boldsymbol{x}'\ CI\ \boldsymbol{x}''$, then we define $\boldsymbol{x}\ CI^*\ \boldsymbol{x}'$, $\boldsymbol{x}'\ CI^*\ \boldsymbol{x}''$, and $\boldsymbol{x}\ CI^*\ \boldsymbol{x}''$. This reflects the property that we can go from $\boldsymbol{x}$ to $\boldsymbol{x}''$ via $\boldsymbol{x}'$. This makes the common input bag relation CI^* an equivalence relation on $C\ell T$.

Definition 10.10 (Common input bag class derivatives). Let $\mathcal{C} = \{CI^1, \ldots, CI^\ell\}$ be the set of all common input bag classes. Let the *transition set* $\mathcal{T}(CI^i)$ of a common input bag class CI^i be the set of all transitions that are used by the closed support T-invariants in CI^i, i.e.,

$$\mathcal{T}(CI^i) = \{t \in T | \exists \boldsymbol{x} \in CI^i : t \in ||\boldsymbol{x}||\}.$$

Let the *place set* $\mathcal{P}(CI^i)$ of common input bag class CI^i be the set of all places are elements of the closed support T-invariants in CI^i, i.e.,

$$\mathcal{P}(CI^i) = \{p \in P | \exists t \in \mathcal{T}(CI^i) : I_p(t) > 0\}.$$

Finally, we say that common input bag classes CI^i and CI^j are *connected* if $\mathcal{P}(CI^i) \bigcap \mathcal{P}(CI^j) \neq \emptyset$.

The common input bag classes partition $C\ell T$: each $\boldsymbol{x} \in C\ell T$ belongs to exactly one common input bag class. Let $\boldsymbol{x} \in C\ell T$ with equivalence class $CI(\boldsymbol{x})$. The partitioning of $C\ell T$ into equivalence classes $\{CI(\boldsymbol{x})\}_{\boldsymbol{x} \in C\ell T}$ induces a partition $\{\mathcal{R}(\mathcal{T}(CI(\boldsymbol{x}))\}_{\boldsymbol{x} \in C\ell T}$ of S into irreducible sets of the Markov chain $\mathbf{Y}$ if and only if all transitions are covered by minimal closed support T-invariants [68]. To this end, note that first $\mathcal{R}(\mathcal{T}(CI(\boldsymbol{x}'))) = \mathcal{R}(\mathcal{T}(CI(\boldsymbol{x}))$ if $CI(\boldsymbol{x}') = CI(\boldsymbol{x})$, and $\mathcal{R}(\mathcal{T}(CI(\boldsymbol{x}')) \cap \mathcal{R}(\mathcal{T}(CI(\boldsymbol{x})) = \emptyset$ if $CI(\boldsymbol{x}') \cap CI(\boldsymbol{x}) = \emptyset$. Second, by definition, the input bags $\boldsymbol{I}(t)$ in a set $\mathcal{R}(\mathcal{T}(CI(\boldsymbol{x}))$ are communicating states. Third, when every transition is covered by a minimal closed support T-invariant, each transition is contained in a set $\mathcal{R}(\mathcal{T}(CI(\boldsymbol{x})) \in S$. Thus, for an $\mathcal{S}\Pi$-net, the structure of the minimal closed support T-invariants implies that the routing chain partitions into $|\mathcal{C}| = \ell$ irreducible sets: $\mathcal{R}(\mathcal{T}(CI^i)), i = 1, \ldots, \ell$. This yields that the global balance equations for the routing chain partition into ℓ independent systems of equations, which all have a unique solution up to a multiplicative constant. This leads to the following theorem.

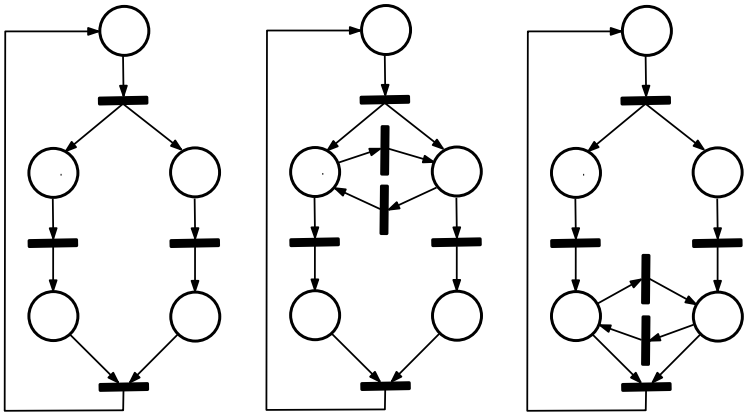
Theorem 10.11. ([68]) For the stochastic Petri net $\mathcal{SPN}$ a positive solution for the routing chain (10.8) exists if and only if $\mathcal{SPN}$ is an $\mathcal{S}\Pi$-net.

In the next corollary, Theorem 10.11 is expanded to the reachability set level. A proof is omitted, as it follows exactly the lines as the proof of Theorem 10.11.

Corollary 10.12. For an $\mathcal{S}\Pi$-net, there is a one-to-one mapping between the partitioning of S into irreducible sets $\{\mathcal{R}(\mathcal{T}(CI(\boldsymbol{x}))\}_{\boldsymbol{x} \in C\ell T}$ that is induced by the partitioning of $C\ell T$ into equivalence classes $\{CI(\boldsymbol{x})\}_{\boldsymbol{x} \in C\ell T}$ and the partitioning of local state spaces $V(\boldsymbol{n})$ into the local irreducible sets $V_i(\boldsymbol{n})$.

The next Theorem shows that an $\mathcal{S}\Pi$-net not only guarantees a positive solution for the global balance equations for the routing chain (10.8), but for live initial markings also for the global balance equations (9.4) for the Markov chain $\mathbf{X}$ of the stochastic Petri net.

Theorem 10.13 ([67])**.** A marked Π-net $\mathcal{PN} = (P, T, I, O, \boldsymbol{m}_0)$ underlying a marked $\mathcal{S}\Pi$-net $(\mathcal{SPN}, \boldsymbol{m}_0)$ has home state $\boldsymbol{m}_0$ and is structurally live.



(a) (b) (c)

Figure 10.2: Petri nets of Remark 10.14.

If the net is covered by P-invariants, it is structurally bounded (Result 9.26). Positive recurrence then follows and thus a positive solution solution summing to unity exists. Furthermore, Theorem 10.13 shows that there exists an initial marking for which the net is live. The proof indicates that if each common input bag is initially marked, the net is live. If it is not the case that each common input bag is initially marked, checking liveness may be cumbersome (see Haddad et al. [270]).

Remark 10.14. When the equilibrium behavior of stochastic Petri nets is of interest, a natural condition is that all transitions are covered by minimal support T-invariants. For bounded nets this condition is necessary for liveness (see Result 9.26). If this condition is not satisfied, there exists a transition, say t_0, that is enabled in a reachable marking $\boldsymbol{m}$, and $\boldsymbol{x}(t_0) = 0$ for all minimal support T-invariants (if t_0 is never enabled, then we can delete t_0 from T). Let t_0 fire in marking $\boldsymbol{m}$. Then there exists no firing sequence from $\boldsymbol{m} - \boldsymbol{I}(t_0) + \boldsymbol{O}(t_0)$ back to $\boldsymbol{m}$ (otherwise t_0 would be contained in a T-invariant). Thus $\boldsymbol{m}$ is a transient state and does not appear in the equilibrium description of the stochastic Petri net. As a consequence, both $\boldsymbol{m}$ and t_0 can be deleted from the equilibrium description of the Petri net.

Observe the Petri nets in figure 10.2a-10.2c, which are not $\mathcal{S}\Pi$-nets. As can be seen from the Petri net of Figure 10.2b, the condition that all transitions are covered by T-invariants is necessary, but not sufficient for liveness of the Petri net. For liveness additional conditions are required.

An $\mathcal{S}\Pi$-net does guarantee structural liveness of the Petri net. As can be seen from Figure 10.2a, and 10.2c, the condition of an $\mathcal{SPN}$ being an $\mathcal{S}\Pi$-net is sufficient, but not necessary. Comparison of Figure 10.2b, and 10.2c, however, shows that the property of liveness is cumbersome since Petri nets that are almost identical may show completely different behavior. Therefore, a characterization of liveness for $\mathcal{S}\Pi$-nets is of interest on its own. □

10.3.2 Group-local-balance and product form

In Section 10.3.1, we have first seen that if GLB holds, a positive solution to the routing chain (10.8) and thus to the local balance equations (10.2) is guaranteed. Second, a positive solution to the routing chain exists if and only if the stochastic Petri net is an $\mathcal{S}\Pi$-net. In this section, we investigate the equivalence of GLB and a product form solution over the places of the Petri net. As can be seen from Corollary 10.6, a positive solution to the routing chain does not yet imply GLB and thus a product from solution. The additional condition to be satisfied is also formulated in this section, of which the structural implications are discussed in Section 10.3.3.

From Corollary 10.6 we obtain the key idea that under GLB the marking independent solution $y(\cdot)$ of the routing chain can be translated into a marking dependent solution with the same properties. This is reflected by the ratio condition (10.5). Also, from the analysis in Section 10.3.1 we know that $x(\boldsymbol{I}(t);\boldsymbol{n}) = y(\boldsymbol{I}(t))$ is a solution to the local balance equations (10.2). For state independent firing rates this leads to the following corollary, which is similar to Theorem 1 of Henderson and Taylor [299].

Corollary 10.15. The equilibrium distribution π of an $\mathcal{SPN}$ with state independent firing rates satisfies GLB if and only if it is an $\mathcal{S}\Pi$-net and a function $\pi_y : \mathcal{M}(\mathcal{SPN},\boldsymbol{m}_0) \to \mathbb{R}^+$ exists such that for all $\boldsymbol{n}+\boldsymbol{I}(t) \in \mathcal{M}(\mathcal{SPN},\boldsymbol{m}_0)$, $t,t' \in T$ with $p(\boldsymbol{I}(t),\boldsymbol{I}(t')) > 0$,

$$\frac{\pi_y(\boldsymbol{n}+\boldsymbol{I}(t))}{\pi_y(\boldsymbol{n}+\boldsymbol{I}(t'))} = \frac{y(\boldsymbol{I}(t))}{y(\boldsymbol{I}(t'))}, \tag{10.9}$$

and $\pi(\boldsymbol{m}) = B\pi_y(\boldsymbol{m})$, $\boldsymbol{m} \in \mathcal{M}(\mathcal{SPN},\boldsymbol{m}_0)$ with $B^{-1} = \sum_{\boldsymbol{m}\in\mathcal{M}(\mathcal{SPN},\boldsymbol{m}_0)} \pi_y(\boldsymbol{m})$ is the unique equilibrium distribution of the Markov chain describing $\mathcal{SPN}$.

Note that Condition (10.9) is a condition on y and *not* on the structure of the Petri net. If a solution $y(\cdot)$ for the routing chain is found, a function $\pi_y(\cdot)$ satisfying (10.9) cannot always be found without additional assumptions on the $\mathcal{SPN}$. Theorem 10.19 below provides a product form solution for π_y under additional conditions on the Petri net. To formulate and understand the structural characterization of the $\mathcal{SPN}$s guaranteeing the ratio condition (10.9), first Lemmas 10.16 and 10.18 and Corollary 10.17 are presented.

Corollary 10.15 implies that the equilibrium distribution π of an $\mathcal{S}\Pi$-net with state independent firing rates satisfies GLB if and only if for an arbitrary reference state $\boldsymbol{m}_0$, and all $\boldsymbol{m} \in \mathcal{M}(\mathcal{SPN},\boldsymbol{m}_0)$

$$\pi(\boldsymbol{m}) = \pi(\boldsymbol{m}_0)\prod_{k=0}^{s}\frac{y(\boldsymbol{I}(t_k))}{y(\boldsymbol{I}(t'_k))}, \tag{10.10}$$

for all firing sequences of the form

$$\begin{aligned}\boldsymbol{m}_0 = \boldsymbol{n}_0 + I(t_0) \to \boldsymbol{n}_0 + I(t'_0) = \boldsymbol{n}_1 + \boldsymbol{I}(t_1) \to \boldsymbol{n}_1 + I(t'_1) = \ldots \to \\ \ldots = \boldsymbol{n}_s + \boldsymbol{I}(t_s) \to \boldsymbol{n}_s + \boldsymbol{I}(t'_s) = \boldsymbol{n}_{s+1} + \boldsymbol{I}(t_{s+1}) = \boldsymbol{m}\end{aligned}$$

This is seen by first observing that for state independent firing rates $x(\boldsymbol{I}(t);\boldsymbol{n}) = y(\boldsymbol{I}(t))$ is a solution of the local balance equations (10.2) and then substituting (10.3) in (10.4) of Theorem 10.5. Applying (10.10) to a cyclic firing sequence, so for $\boldsymbol{m}_0 = \boldsymbol{m}$, yields the following lemma.

Lemma 10.16. The equilibrium distribution π of an $\mathcal{S}\Pi$-net with state independent firing rates (10.6) satisfies GLB if and only if for each T-invariant $\boldsymbol{x} = (x_1, \ldots, x_M)$

$$\prod_{t=1}^{M} \left(\frac{y(\boldsymbol{I}(t))}{y(\boldsymbol{O}(t))} \right)^{x_t} = 1. \tag{10.11}$$

In Section 10.3.3, we will investigate the structural Petri net conditions that Lemma 10.16 imposes. First, we will use Lemma 10.16 in showing that a solution π_y satisfying the ratio condition (10.9) must be a product form over the places of the network.

Following Coleman et al. [130], we introduce the row vector $\boldsymbol{C}(y)$, defined as $\boldsymbol{C}(y)_t = \log\big(y(\boldsymbol{I}(t))/y(\boldsymbol{O}(t))\big)$. As $y(\cdot)$ is determined up to a multiplicative constant, and $\boldsymbol{C}(y)$ is determined by the ratios of y's, the vector $\boldsymbol{C}(y)$ is unique, so that can safely be denoted by $\boldsymbol{C}$. Taking logarithms on both sides in equation (10.11), Lemma 10.16 can now be reformulated as follows.

Corollary 10.17. The equilibrium distribution π of an $\mathcal{S}\Pi$-net with firing rates (10.6) satisfies GLB if and only if $\boldsymbol{C}\boldsymbol{x} = 0$ for every T-invariant $\boldsymbol{x}$.

Lemma 10.18 ([129])**.** The following statements are equivalent:

(i) $\boldsymbol{C}\boldsymbol{x} = 0$ for each T-invariant $\boldsymbol{x}$.

(ii) $Rank[\boldsymbol{A}] = Rank[\boldsymbol{A}|\boldsymbol{C}]$, where $[\boldsymbol{A}|\boldsymbol{C}]$ is the matrix augmented with row $\boldsymbol{C}$.

(iii) Equation $\boldsymbol{z}\boldsymbol{A} = \boldsymbol{C}$ has a solution $\boldsymbol{z}$.

Proof. The lemma was stated without proof in [129]. For completeness, it is provided here.

(i) ⇒ (ii) Assume (i) is true. This is, for each $\boldsymbol{x}$ such that $\boldsymbol{A}\boldsymbol{x} = 0$, also $\boldsymbol{C}\boldsymbol{x} = 0$. This implies that the kernel of $\boldsymbol{A}$ is a subspace of the kernel of $[\boldsymbol{A}|\boldsymbol{C}]$, which induces $dim(ker(\boldsymbol{A})) \leq dim(ker([\boldsymbol{A}|\boldsymbol{C}])$. Hence, $rank(\boldsymbol{A}) \geq Rank([\boldsymbol{A}|\boldsymbol{C}])$. Of course, since A is a submatrix of $[\boldsymbol{A}|\boldsymbol{C}]$, also $rank(\boldsymbol{A}) \leq rank([\boldsymbol{A}|\boldsymbol{C}])$. Combining these relations yields $Rank(\boldsymbol{A}) = Rank([\boldsymbol{A}|\boldsymbol{C}])$.

(ii) ⇒ (iii) $Rank(\boldsymbol{A}) = Rank([\boldsymbol{A}|\boldsymbol{C}])$ implies that the row vector $\boldsymbol{C}$ can be written as a linear combination of the rows of $\boldsymbol{A}$, i.e., $\boldsymbol{z}\boldsymbol{A} = \boldsymbol{C}$ has a solution.

(iii) ⇒ (i) $\boldsymbol{z}\boldsymbol{A} = \boldsymbol{C}$ has a solution means that the row vector $\boldsymbol{C}$ can be written as a linear combination of the rows of $\boldsymbol{A}$. For a T-invariant $\boldsymbol{A}\boldsymbol{x} = 0$. Combining these statements implies $\boldsymbol{C}\boldsymbol{x} = 0$. □

Using Lemma 10.18, the following key-result identifies the equivalence between GLB and a product form solution over the places of the network. The solution z of the condition 3. is used to express the product form. Section 10.3.3 investigates the intuition behind this theorem and provides an explanation in terms of T-invariants.

Theorem 10.19. Consider an $\mathcal{SPN}$ with state independent firing rates (10.6). The equilibrium distribution π satisfies GLB if and only if the $\mathcal{SPN}$ is an $\mathcal{S}\Pi$-net, $\boldsymbol{z}\boldsymbol{A} = \boldsymbol{C}$ has a solution and π is a product form over the places of the network

$$\pi_y(\boldsymbol{m}) = \prod_{p=1}^{N} (f_p)^{m_p}, \qquad \boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0), \tag{10.12}$$

where $f_p = e^{-z_p}$ and $\pi(\boldsymbol{m}) = B\pi_y(\boldsymbol{m})$ with $B^{-1} = \sum_{\boldsymbol{m}\in\mathcal{M}(\mathcal{SPN},\boldsymbol{m}_0)} \pi_y(\boldsymbol{m})$.

Proof. Under GLB, by Corollary 10.17, $\boldsymbol{C}\boldsymbol{x} = 0$ for each minimal support T-invariant. This implies by lemma 10.18 that the equation $\boldsymbol{z}\boldsymbol{A} = \boldsymbol{C}$ has a solution. Thus we obtain for each transition $t \in T$

$$\sum_{p=1}^{N} z_p A(p,t) = \log\left(\frac{y(\boldsymbol{I}(t))}{y(\boldsymbol{O}(t))}\right).$$

Taking exponentials gives

$$\prod_{p=1}^{N} e^{z_p A(p,t)} = \left(\frac{y(\boldsymbol{I}(t))}{y(\boldsymbol{O}(t))}\right).$$

By Corollary 10.15, we then have for all $\boldsymbol{n} + \boldsymbol{I}(t) \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)$, $t, t' \in T$ with $p(\boldsymbol{I}(t), \boldsymbol{I}(t')) > 0$

$$\frac{\pi_y(\boldsymbol{n}+\boldsymbol{I}(t))}{\pi_y(\boldsymbol{n}+\boldsymbol{I}(t'))} = \frac{y(\boldsymbol{I}(t))}{y(\boldsymbol{I}(t'))} = \prod_{p=1}^{N} e^{z_p A(p,t)}.$$

By (10.10), for all markings $\boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)$, $\pi(\boldsymbol{m})$ can be expressed in terms of the reference state $\boldsymbol{m}_0$

$$\begin{aligned}
\pi(\boldsymbol{m}) &= \pi(\boldsymbol{m}_0) \prod_{k=0}^{s} \prod_{p=1}^{N} e^{z_i A(i,t_k)} = \pi(\boldsymbol{m}_0) \prod_{p=1}^{N} e^{z_p(m_0(p)-m(p))} \\
&= \pi(\boldsymbol{m}_0) \left\{ \prod_{p=1}^{N} e^{z_p m_0(p)} \right\} \left\{ \prod_{p=1}^{N} e^{-z_p m(p)} \right\} = B \prod_{p=1}^{N} (f_p)^{m(p)} = B\pi_y(\boldsymbol{m}).
\end{aligned}$$

Conversely, if an $\mathcal{S}\Pi$-net has an equilibrium distribution $\pi(\boldsymbol{m}) = B\prod_{p=1}^{N} f_p^{m(p)}$, then GLB is satisfied, since for an $\mathcal{S}\Pi$-net the GLB equations (10.1) reduce to

$$\pi(\boldsymbol{n}+\boldsymbol{I}(t)) \sum_{t'\in T} q(\boldsymbol{I}(t), \boldsymbol{I}(t'); \boldsymbol{n}) = \sum_{t'\in T} \pi(\boldsymbol{n}+\boldsymbol{I}(t')) q(\boldsymbol{I}(t'), \boldsymbol{I}(t); \boldsymbol{n}), \tag{10.13}$$

for all $\boldsymbol{n}, \boldsymbol{I}(t)$ such that $\boldsymbol{n} + \boldsymbol{I}(t) \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)$. Substituting (10.12) into (10.13) and dividing by $B\prod_{p=1}^{N} f_p^{n_p}$ yields

$$\prod_{p=1}^{N} f_p^{(I_p(t))} \sum_{t'\in T} \mu(t) p(\boldsymbol{I}(t), \boldsymbol{I}(t')) = \sum_{t'\in T} \prod_{p=1}^{N} f_p^{(I_p(t'))} \mu(t') p(\boldsymbol{I}(t'), \boldsymbol{I}(t)).$$

We recognize the routing chain equations (10.8). The solution $y(\cdot)$ to the routing chain is unique. So for the GLB-equations to be verified, it remains to show that, for all $t \in T$

$$\prod_{p=1}^{N} f_p^{(\boldsymbol{I}_p(t))} = y(\boldsymbol{I}(t)). \tag{10.14}$$

To this end, note that by the definition of the f_p's

$$\begin{aligned} \log\left(\frac{y(\boldsymbol{I}(t))}{y(\boldsymbol{O}(t))}\right) &= \sum_{p=1}^{N} A(p,t) z_p = \sum_{p=1}^{N} \boldsymbol{I}_p(t)\log(f_p) - \boldsymbol{O}_p(t)\log(f_p) \\ &= \sum_{p=1}^{N} \log\left(\frac{f_p^{(\boldsymbol{I}_p(t))}}{f_p^{(\boldsymbol{O}_p(t))}}\right), \end{aligned}$$

and thus

$$\frac{y(\boldsymbol{I}(t))}{y(\boldsymbol{O}(t))} = \prod_{p=1}^{N} \left(\frac{f_p^{(\boldsymbol{I}_p(t))}}{f_p^{(\boldsymbol{O}_p(t))}}\right),$$

which shows that (10.14) is satisfied. □

Under the condition that a solution to the routing chain exists, equivalence of condition (ii) of Lemma (10.18) and product form π_y satisfying (10.9), was obtained by Coleman et al. [130]. The solution z of the alternative condition (iii) was used to express the explicit solution of the product form. The contribution of Theorem 10.19 is the explicit relation between GLB and product form.

Theorem 10.19 characterizes product forms for $\mathcal{SPN}$s based on the incidence matrix. The product form (10.12) is of the Jackson-type since it is a product over the places similar to the result of Jackson [330]. Note that Petri nets are substantially more complex than Jackson networks. The product form distribution (10.12) contains one term for each token in the Petri net. Therefore, under GLB the only dependence between tokens lies in the normalising constant, as is the case in closed Jackson networks. Observe that Theorem 10.19 does not state that an arbitrary $\mathcal{SPN}$ with product form equilibrium distribution satisfies GLB.

Remark 10.20. Each T-invariant can be written as a linear combination of minimal support T-invariants (result 9.27). Therefore, it can readily be seen that in Lemma 10.16, Corollary 10.17 and Lemma 10.18 the statement 'for each T-invariant', can be replaced by 'for each minimal support T-invariant'. This observation will be convenient when studying the structural implications of the results presented in this section.

10.3.3 Structural implications of product form $\mathcal{SPN}$s

In this section, we study the structural implication of Theorem 10.19 on the Petri net. The condition $Rank[\boldsymbol{A}] = Rank[\boldsymbol{A}|\boldsymbol{C}]$ was presented in Coleman et al. [130] as a necessary and sufficient condition for product form. Three comments can be placed regarding their results: (1) they assumed that a solution of the routing chain exists, (2) the condition $Rank[\boldsymbol{A}] = Rank[\boldsymbol{A}|\boldsymbol{C}]$ generally depends on the numerical values of the transition rates, and (3) $Rank[\boldsymbol{A}] = Rank[\boldsymbol{A}|\boldsymbol{C}]$ is a technical condition without intuitive interpretation.

The first comment is addressed in Theorem 10.11; for a solution of the routing chain to exist the Petri net must be an $\mathcal{S}\Pi$-net. The second comment was already observed by Coleman et al. [130], where it is shown that in some cases conditions on the numerical values of the firing rates must be imposed and in some cases not. To this end, Haddad et al. [270] introduced $\mathcal{S}\Pi^2$-nets, a subclass of $\mathcal{S}\Pi$-nets that have product form irrespective of the numerical values of the firing rates. Mairesse et al. [406] relate the Deficiency Zero Theorem of Feinberg [211], developed for chemical reaction networks, to product form results for stochastic Petri nets. They show that the concept of $\mathcal{S}\Pi^2$-nets coincides with $\mathcal{S}\Pi$-nets that have 'deficiency zero'. However, neither the characterization of $\mathcal{S}\Pi^2$-nets or deficiency-zero $\mathcal{S}\Pi$-nets do intuitively explain why no restrictions on the numerical values of the firing rates are imposed. The structural implications of the product form results of Theorem 10.19, are based on the *minimal support* T-invariants (see Remark 10.20). First, we will show that $\mathcal{S}\Pi$-nets in which all minimal support T-invariants are minimal *closed* support T-invariants have product form without additional conditions on the firing rates. Second, we will show that this characterization exactly corresponds to the definition of $\mathcal{S}\Pi^2$-nets provided by Haddad et al. [270] and deficiency-zero $\mathcal{S}\Pi$-nets provided by Mairesse et al. [406]. Third, via this characterization in terms of the minimal support T-invariants we are able to provide an explanation in terms of T-invariants of the condition $Rank[\boldsymbol{A}] = Rank[\boldsymbol{A}|\boldsymbol{C}]$ of the $\mathcal{SPN}$. The condition is shown to be required only for $\mathcal{S}\Pi$-nets that are not $\mathcal{S}\Pi^2$-nets.

Theorem 10.21. For an $\mathcal{SPN}$, (10.11) is satisfied for each minimal *closed* support T-invariant $\boldsymbol{x}$. For an $\mathcal{S}\Pi$-net in which each minimal support T-invariant is a minimal *closed* support T-invariant, the equivalent conditions (i)-(iii) of Lemma 10.18 are satisfied.

Proof. The firing sequence of a minimal closed support T-invariant is linear (see Result 9.29). Thus, $x_t \leq 1,\ t = 1, \ldots, T$, and within this T-invariant every output bag is an input bag of a unique next transition. Therefore, in (10.11) the denominator of each fraction $y(\boldsymbol{I}(t))/y(\boldsymbol{O}(t))$ is canceled by the numerator of the fraction of the subsequent transition in this T-invariant. As a consequence, conditions (i)-(iii) of Lemma 10.18 are satisfied irrespective of of the numerical values of the firing rates. □

By means of Theorem 10.21, in the case that there exists a minimal T-invariant that is *not* closed, additional conditions are required on the numerical values of

the firing rates to ensure a product form solution. Below, we will provide an intuitive explanation of these additional conditions. First, the definition of $\mathcal{S}\Pi^2$-nets, as introduced by Haddad et al. [270], is presented.

Definition 10.22 ($\mathcal{S}\Pi^2$-net [270])**.** A Π^2-net is a Π-net such that for every $\boldsymbol{r} \in \mathcal{R}(\mathcal{T})$, there is an $\boldsymbol{a_r} \in \mathbb{Q}^N$ such that

$$\boldsymbol{a_r A} = \boldsymbol{b_r},$$

in which for $t = 1, \ldots, N$

$$\boldsymbol{b_r}(t) = \begin{cases} -1 & \text{, if } \boldsymbol{r} = \boldsymbol{I}(t), \\ 1 & \text{, if } \boldsymbol{r} = \boldsymbol{O}(t), \\ 0 & \text{, otherwise.} \end{cases}$$

An $\mathcal{S}\Pi^2$-net is a stochastic Π^2-net.

Although not defined as such by Haddad et al. [270], and not recognized before, as is shown in the next theorem, the characterization of an $\mathcal{S}\Pi^2$-net can be provided via the minimal support T-invariants of the $\mathcal{S}\Pi$-net.

Theorem 10.23. An $\mathcal{S}\Pi$-net is an $\mathcal{S}\Pi^2$-net if and only if *all* its minimal support T-invariants are minimal *closed* support T-invariants.

Proof. Consider an $\mathcal{S}\Pi$-net. We must show that $\boldsymbol{a_r A} = \boldsymbol{b_r}$ has a solution if and only if all minimal support T-invariants are minimal closed support T-invariants. First observe that $\boldsymbol{a_r A} = \boldsymbol{b_r}$ has a solution if and only if the row vector $\boldsymbol{b_r}$ is a linear combination of the rows of $\boldsymbol{A}$, i.e., $\boldsymbol{b_r x} = 0$ for every $\boldsymbol{x}$ such that $\boldsymbol{Ax} = 0$, that is $\boldsymbol{b_r x} = 0$ for all T-invariants. Second, if a solution $\boldsymbol{a_r}$ exists, it is rational since $\boldsymbol{A}$ is an integer matrix and $\boldsymbol{b_r}$ an integer vector.

Now, assume that all minimal support T-invariants are minimal closed support. Consider a minimal closed support T-invariant $\boldsymbol{x}$ and a bag $\boldsymbol{r} \in \mathcal{R}(T)$ with $\boldsymbol{O}(t_i) = \boldsymbol{I}(t_j)$, then $\boldsymbol{b_r x} = x_{t_i} - x_{t_j}$, since the firing sequence of $\boldsymbol{x}$ is linear (Result 9.29). Either $\boldsymbol{r}$ is both an input bag and an output bag of transitions in the firing sequence of $\boldsymbol{x}$ (i.e., $x_{t_i} = x_{t_j} = 1$), or $\boldsymbol{r}$ is neither an input bag nor an output bag of any transition in the firing sequence of $\boldsymbol{x}$ (i.e., $x_{t_i} = x_{t_j} = 0$). By assumption all minimal support T-invariants are minimal closed support, which completes the first part of the proof.

Conversely, if there is a minimal support T-invariant $\boldsymbol{x}$ of which the support is not closed, then $\exists \boldsymbol{r} \in \mathcal{R}(T), t \in \|\boldsymbol{x}\|$, such that $\boldsymbol{b}$ is the output of t, but there is no $t' \in \|\boldsymbol{x}\|$ such that $\boldsymbol{r}$ is the input bag of t'. For such $\boldsymbol{x}$ we have $\boldsymbol{b_r x} \neq 0$ and this completes the proof of the second part. □

Corollary 10.24. For an $\mathcal{S}\Pi^2$-net the equivalent conditions (i)–(iii) of Lemma 10.18 are satisfied irrespective of the firing rates. Therefore, GLB and a product form solution of the form (10.12) can be verified without checking one of these conditions.

Proof. By Theorem 10.21 and Theorem 10.23, for an $\mathcal{S}\Pi^2$-net the equivalent conditions (i)–(iii) of Lemma 10.18 are satisfied irrespective of the transition rates. Applying Theorem 10.19 concludes the proof. □

Now, we give the definition of the deficiency of a Petri net. Mairesse et al. [406] show that $\mathcal{S}\Pi$-nets that have deficiency zero have a product form equilibrium distribution irrespective of the numerical values of the transition rates. They also observe that the class of zero-deficiency $\mathcal{S}\Pi$-nets coincides with that of $\mathcal{S}\Pi^2$-nets.

Definition 10.25 (Deficiency [406])**.** The deficiency δ of a Petri net $\mathcal{PN}$ is:

$$\delta = |\mathcal{R}(\mathcal{T})| - \ell - rank(\boldsymbol{A}),$$

where $|\mathcal{R}(\mathcal{T})|$ represents the number of bags $\boldsymbol{r} \in \mathcal{R}(\mathcal{T})$ and ℓ is the number of common input bag classes of $\mathcal{PN}$.

Lemma 10.26 ([406])**.** Consider an $\mathcal{S}\Pi$-net $\mathcal{SPN}$. $\mathcal{SPN}$ is an $\mathcal{S}\Pi^2$-net if and only if it has deficiency $\delta = 0$.

Theorem 10.23 and Lemma 10.26 imply that for $\mathcal{S}\Pi$-net deficiency zero is a property that can also be identified via its minimal support *T*-invariants. Deficiency is directly related to the number of linearly independent minimal *non*-closed support *T*-invariants.

To conclude, Theorem 10.19 states that the equilibrium distribution of an $\mathcal{S}\Pi$-net is characterized by the solution of the routing chain $y(\cdot)$, characterized by the probability flow through classes of minimal closed support *T*-invariants. In $\mathcal{S}\Pi$-nets, all transitions are covered by minimal closed support *T*-invariants. Therefore, every minimal support *T*-invariant that is not closed support is built up by transitions of different minimal closed support *T*-invariants. The conditions (i)-(iii) of Lemma 10.18 imply that the total probability flow through a minimal non-closed support *T*-invariant should equal to the probability flow imposed by the minimal closed support *T*-invariants. Examples 10.29 and 10.30 in the next subsection will provide an illustration.

From the results presented above, it is clear that characterization of product form results for $\mathcal{SPN}$s with transition rates (10.6) can be done at the structural level. The steps that have to be performed to this end are summarized in the following algorithm.

Algorithm 10.27 (Structural characterization of product form)**.**

Step 1. *Obtain the incidence matrix $\boldsymbol{A}$ of the $\mathcal{SPN}$ and compute the minimal support T-invariants $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^h$ and the minimal support P-invariants $\boldsymbol{y}^1, \ldots, \boldsymbol{y}^j$.*

Step 2. *Obtain the minimal closed support T-invariants from the minimal support T-invariants, and renumber the T-invariants such that $\{\boldsymbol{x}^1, \ldots, \boldsymbol{x}^k\}$ is the set of minimal closed support T-invariants ($k \leq h$).*

Step 3. *Verify that all transitions are covered by minimal closed support T-invariants and minimal support P-invariants. If not: stop, we cannot conclude that the $\mathcal{SPN}$ has a product form equilibrium distribution, else: go to step 4.*

Step 4. *Determine from* $\{\boldsymbol{x}^1,\ldots,\boldsymbol{x}^k\}$ *the set of common input bag classes* $\mathcal{C}=\{CI^1,\ldots,CI^\ell\}$. *Compute per common input bag class* i *the solution to the routing chain* $y^i(\cdot)$. *If all minimal support T-invariants are minimal closed support T-invariants, i.e.,* $k=h$, *then proceed to step 6, else go to step 5.*

Step 5. *Determine* $\boldsymbol{C}$ *and verify that* $\boldsymbol{C}\boldsymbol{x}^i=0$, *for the minimal non-closed support T-invariants* $\boldsymbol{x}^{k+1},\ldots,\boldsymbol{x}^h$. *If not: stop, the* $\mathcal{SPN}$ *does not have a product form equilibrium distribution, else go to step 6.*

Step 6. *Solve* $\mathbf{z}\boldsymbol{A}=\boldsymbol{C}$. *The equilibrium distribution is* $\pi(\boldsymbol{m})=B\pi_y(\boldsymbol{m})$ *with* π_y *given in (10.12).*

10.4 Examples

This section presents some examples illustrating the structural characterization of product form presented above. First, in Example 10.28 we present an example of an $\mathcal{S}\Pi^2$-net. Examples 10.29 and 10.30 present $\mathcal{S}\Pi$-nets that are not $\mathcal{S}\Pi^2$-nets, which means that they have a product form equilibrium distribution only for specific choices of the firing rates. Finally, in Example 10.31, we illustrate the importance of the boundedness assumption, by presenting a net that may not possess an equilibrium distribution, due to a possibly unbounded number of tokens. Examples 10.28, 10.29 and 10.31 are obtained from [66].

Example 10.28. Consider the $\mathcal{SPN}$ depicted in Figure 10.3a and execute the steps of the algorithm of Section 10.3.3.

Step 1–3. From the incidence matrix

$$\boldsymbol{A}=\begin{pmatrix} -1 & -1 & 1 & 0 & 0\\ 1 & 0 & -1 & 1 & 0\\ 2 & 1 & -2 & 2 & -1\\ 0 & 1 & 0 & 0 & -1\\ 0 & 0 & 0 & -1 & 1 \end{pmatrix},$$

we obtain that this net has two minimal support T-invariants $\boldsymbol{x}^1=(10100)$, $\boldsymbol{x}^2=(01111)$, which are both minimal closed support T-invariants, and two minimal support P-invariants $\boldsymbol{y}^1=(11011)$, $\boldsymbol{y}^2=(20112)$. $\mathcal{SPN}$ is covered by both minimal support T-invariants and P-invariants.

Step 4. Since the T-invariants share $\boldsymbol{I}(t_1)$ they are in common input bag relation, which implies that the routing chain has one irreducible set:

$$S=\{\boldsymbol{I}(t_1),\boldsymbol{I}(t_3),\boldsymbol{I}(t_4),\boldsymbol{I}(t_5)\}\ \ (\boldsymbol{I}(t_1)=\boldsymbol{I}(t_2)).$$

Amalgamate transition t_1 and t_2 into a single transition t_{12} with $\mu(t_{12})=\mu(t_1)+\mu(t_2)$, $p(\boldsymbol{I}(t_1),\boldsymbol{O}(t_1))=\mu(t_1)/\mu(t_{12})$ and $p(\boldsymbol{I}(t_1),\boldsymbol{O}(t_2))=\mu(t_2)/\mu(t_{12})$. The solu-

tion of the routing chain is (up to normalization):

$$y(\boldsymbol{I}(t_1))\mu(t_{12}) = y(\boldsymbol{I}(t_3))\mu(t_3) = 1,$$
$$y(\boldsymbol{I}(t_4))\mu(t_4) = y(\boldsymbol{I}(t_5))\mu(t_5) = p(\boldsymbol{I}(t_1), \boldsymbol{O}(t_2)).$$

The $\mathcal{SPN}$ is an $\mathcal{S}\Pi^2$-net, so we may proceed to step 6.

Step 6. The vector $\boldsymbol{C}$ is obtained from the solution of the routing chain:

$$\boldsymbol{C} = \left(\log\left[\frac{\mu(t_3)}{\mu(t_{12})}\right], \log\left[\frac{\mu(t_5)}{\mu(t_2)}\right], \log\left[\frac{\mu(t_{12})}{\mu(t_3)}\right], \right.$$
$$\left. \log\left[\frac{\mu(t_2)\mu(t_3)}{\mu(t_{12})\mu(t_4)}\right], \log\left[\frac{\mu(t_4)}{\mu(t_5)}\right] \right).$$

A solution $\boldsymbol{z}$ of $\boldsymbol{zA} = \boldsymbol{C}$ is:

$$z_1 = 0,\ z_2 = \log\left(\frac{\mu(t_3)}{\mu(t_{12})}\right),\ z_3 = 0,\ z_4 = \log\left(\frac{\mu(t_5)}{\mu(t_2)}\right),\ z_5 = \log\left(\frac{\mu(t_4)}{\mu(t_2)}\right),$$

and the equilibrium distribution is

$$\pi(\boldsymbol{m}) = B\left(\frac{\mu(t_{12})}{\mu(t_3)}\right)^{m(2)}\left(\frac{\mu(t_2)}{\mu(t_5)}\right)^{m(4)}\left(\frac{\mu(t_2)}{\mu(t_4)}\right)^{m(5)},$$

for any marking $\boldsymbol{m}$ in the reachability set

$$\mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0) = \{\boldsymbol{m} : \boldsymbol{y}^1(\boldsymbol{m} - \boldsymbol{m}_0) = 0,\ \boldsymbol{y}^2(\boldsymbol{m} - \boldsymbol{m}_0) = 0\},$$

where $\boldsymbol{y}^1 = (11011)$, $\boldsymbol{y}^2 = (20112)$ are the two minimal support P-invariants of the net. □

Example 10.29. Consider the $\mathcal{SPN}$ depicted in Figure 10.3b. This is an example of an $\mathcal{S}\Pi$-net which is not an $\mathcal{S}\Pi^2$-net so that additional conditions on the firing rates have to be satisfied.

Step 1–3. This $\mathcal{SPN}$ has incidence matrix

$$\boldsymbol{A} = \begin{pmatrix} -1 & 1 & -2 & 2 \\ 1 & -1 & 2 & -2 \end{pmatrix}.$$

Observe that each transition is covered by the minimal closed support T-invariants $\boldsymbol{x}^1 = (1100)$, $\boldsymbol{x}^2 = (0011)$, but that $\boldsymbol{x}^3 = (2001)$ and $\boldsymbol{x}^4 = (0210)$ are also minimal support T-invariants that do not have closed support. The $\mathcal{SPN}$ is covered by its one minimal support P-invariant $\boldsymbol{y}^1 = (11)$.

Step 4. The routing chain has two irreducible sets $S(\boldsymbol{x}^1) = \{\boldsymbol{I}(t_1), \boldsymbol{I}(t_2)\}$, and $S(\boldsymbol{x}^2) = \{\boldsymbol{I}(t_3), \boldsymbol{I}(t_4)\}$. The solution of the routing chain is:

$$\frac{y^1(\boldsymbol{I}(t_2))}{y^1(\boldsymbol{I}(t_1))} = \frac{\mu(t_1)}{\mu(t_2)},\quad \frac{y^2(\boldsymbol{I}(t_4))}{y^2(\boldsymbol{I}(t_3))} = \frac{\mu(t_3)}{\mu(t_4)},$$

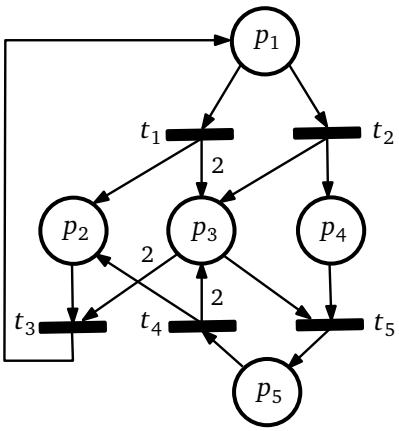
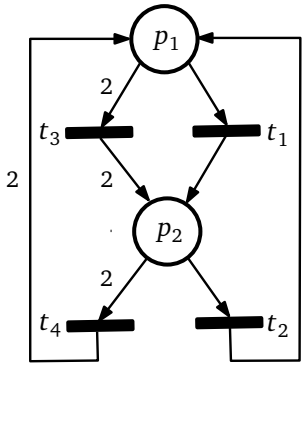


Figure 10.3: (a) $\mathcal{SPN}$ of Example 10.28



(b) $\mathcal{SPN}$ of Example 10.29.

with corresponding vector $\boldsymbol{C}$

$$\boldsymbol{C} = \left(\log \left[\frac{\mu(t_2)}{\mu(t_1)} \right], \log \left[\frac{\mu(t_1)}{\mu(t_2)} \right], \log \left[\frac{\mu(t_4)}{\mu(t_3)} \right], \log \left[\frac{\mu(t_3)}{\mu(t_4)} \right] \right).$$

Step 5. $\boldsymbol{C}\boldsymbol{x}^i = 0$ for the minimal non-closed support T-invariants $\boldsymbol{x}^3 = (2001)$ and $\boldsymbol{x}^4 = (0210)$, if $2C_1 + C_4 = 0$ and $2C_2 + C_3 = 0$, thus if

$$\left(\frac{\mu(t_2)}{\mu(t_1)} \right)^2 = \left(\frac{\mu(t_4)}{\mu(t_3)} \right). \tag{10.15}$$

Step 6. If (10.15) is satisfied, this $\mathcal{SPN}$ has an equilibrium distribution

$$\pi(\boldsymbol{m}) = B \left(\frac{\mu(t_2)}{\mu(t_1)} \right)^{\boldsymbol{m}(1)}.$$

for any marking $\boldsymbol{m}$ in the reachability set

$$\mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0) = \{\boldsymbol{m} : \boldsymbol{m}(1) + \boldsymbol{m}(2) = \boldsymbol{m}_0(1) + \boldsymbol{m}_0(2)\}.$$

This example provides insight in the intuition for the conditions of Lemma 10.18. As can be seen from Figure 10.3b, there are two possibilities for the movement of two tokens from place 1 to place 2. In the first case (via t_1) the tokens jump one after the other, in the second case (via t_3) the tokens jump simultaneously. The probability flow for these two possibilities must be the same. This is reflected in the condition (10.15) on the firing rates: two transitions with rate $\mu(t_1)$ must be proportional to one transition at rate $\mu(t_3)$. □

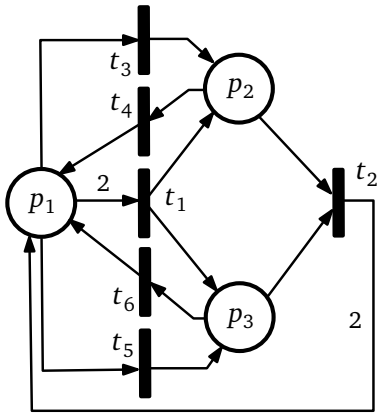
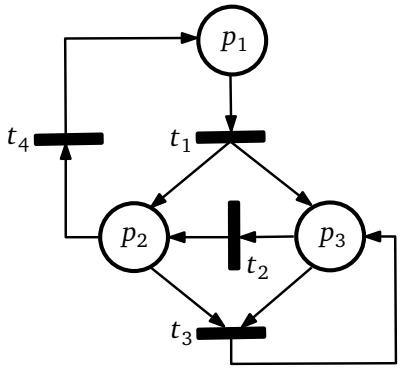


Figure 10.4: (a) $\mathcal{SPN}$ of Example 10.30



(b) $\mathcal{SPN}$ of Example 10.31.

Example 10.30. Consider the $\mathcal{SPN}$ of Figure 10.4a. This example indicates that minimal *non-closed* support T-invariants can also exist in $\mathcal{S}\Pi$-nets where in the minimal support T-invariants no transition fires more than once, i.e., $\boldsymbol{x}_t \leq 1, \forall t \in T$ is not sufficient for a T-invariant to be closed support.

Step 1–3. The minimal closed support T-invariants of this net are $\boldsymbol{x}^1 = (110000)$, $\boldsymbol{x}^2 = (001100)$ and $\boldsymbol{x}^3 = (000011)$ and the minimal non-closed support T-invariants $\boldsymbol{x}^4 = (100101)$ and $\boldsymbol{x}^5 = (011010)$. $\mathcal{SPN}$ is covered by its one minimal support P-invariant $\boldsymbol{y}^1 = (111)$.

Step 4–6. This $\mathcal{SPN}$ has a product form equilibrium distribution if $C_1 = C_4 + C_6$ and $C_2 = C_3 + C_5$, so if

$$\frac{\mu(t_2)}{\mu(t_1)} = \frac{\mu(t_3)}{\mu(t_4)} \frac{\mu(t_5)}{\mu(t_6)}.$$

□

Example 10.31. Consider the $\mathcal{SPN}$ of Figure 10.4b.

Step 1-3. The net has one T-invariant $\boldsymbol{x} = (1111)$ covering all transitions, and $\boldsymbol{x}$ has closed support. It has no P-invariants.

Note that without additional conditions the algorithm stops here. Yet we proceed to provide an illustration of such conditions that prevents the creation of an unbounded number of tokens.

Step 4. The solution of the routing chain is (up to a multiplicative constant)

$$y(\boldsymbol{I}(t_1)) = 1/\mu(t_1), \;\; y(\boldsymbol{I}(t_2)) = 1/\mu(t_2), \;\; y(\boldsymbol{I}(t_3)) = 1/\mu(t_3), \;\; y(\boldsymbol{I}(t_4)) = 1/\mu(t_4),$$

Step 6. The $\mathcal{SPN}$ has an invariant measure

$$\pi_y(\boldsymbol{m}) = \left(\frac{\mu(t_2)\mu(t_4)}{\mu(t_1)\mu(t_3)}\right)^{\boldsymbol{m}(1)} \left(\frac{\mu(t_2)}{\mu(t_3)}\right)^{\boldsymbol{m}(2)} \left(\frac{\mu(t_4)}{\mu(t_3)}\right)^{\boldsymbol{m}(3)}.$$

From Figure 10.4b we can see that the number of tokens in the net is unbounded (repetitive firing of transitions t_1 and t_4 increases the number of tokens by 1), but that for every marking a firing sequence to $\boldsymbol{m}_0 = (100)$ exists. Under the *additional* conditions $\mu(t_2)\mu(t_4) < \mu(t_1)\mu(t_3)$, $\mu(t_2) < \mu(t_3)$, $\mu(t_4) < \mu(t_3)$ the $\mathcal{SPN}$ has an equilibrium distribution

$$\pi(\boldsymbol{m}) = B\pi_y(\boldsymbol{m}), \quad \boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0) = \mathbb{N}_0^3 \setminus \{0\}. \qquad \square$$

CHAPTER 11

Structural Decomposition via Conflict Places

11.1 Introduction

The analysis of Chapter 10 enables us to formulate a decomposition result. This result uses the T- and P-invariants to decompose an $\mathcal{SPN}$ in subnets, consisting of one or more common input bag classes (see Definition 10.9). It is a generalization of the decomposition result formulated by Frosch and Natarajan [223] for Closed Synchronized Systems of Stochastic Sequential Processes ($\mathcal{CS}$) that consist of state machines (see Definition 9.11) connected by so-called buffer places. A formal definition of a $\mathcal{CS}$ is given below in Definition 11.11. By removing these buffer places from the network, the equilibrium (product form) distribution of a $\mathcal{CS}$ is shown to be a product over the product form equilibrium distributions of the separate state machines. As such, this chapter generalizes the results of Frosch and Natarajan to decomposition results for product form $\mathcal{S}\Pi$-nets.

We will formulate sufficient conditions for decomposition of an arbitrary $\mathcal{S}\Pi$-net into subnetworks so that the equilibrium distribution is a product over the invariant measures of the subnetworks defined by common input bag classes. The decomposition is based on conflict places, the generalization of buffer places.

The chapter is organized as follows. First, in Section 11.2, we define three different place sets: the sufficient place set, the surplus place set, and the conflict place set. It will be described how to obtain these place sets from the P- and T-invariants of a Petri net. Section 11.3 presents the decomposition result and is ended with an algorithm by which all possible decompositions of a product form stochastic Petri can be generated. Section 11.4 illustrates the decomposition result and the algorithm along several examples.

11.2 Sufficient, surplus and conflict place sets

The *sufficient place set* was introduced by Florin and Natkin [216]. The places not contained in the sufficient place set will be the places at which we decompose the $\mathcal{SPN}$. We define this complementary set of places as the *surplus place set*.

Definition 11.1 (Sufficient place set - Surplus place set). A subset of places $\mathcal{P}^{suf} \subseteq P$ is a *sufficient place set* if the marking of each place in $\mathcal{P}^{suf}$ provides sufficient information to define uniquely the marking of all places. A subset of places $\mathcal{P}^{sur} \subseteq P$ is a *surplus place set* if the subset of places $P \setminus \mathcal{P}^{sur}$ is a sufficient place set. A place contained in a surplus place set will be referred to as a *surplus place*.

Lemma 11.2. Consider a structurally live and structurally bounded Petri net. A set of places $\mathcal{P} \subseteq P$ is a sufficient place set if and only if all the rows of $\boldsymbol{A}$ can be written as linear combinations of the rows of $\boldsymbol{A}$ corresponding to places in $\mathcal{P}$, i.e., for all $j \in P$,

$$\boldsymbol{A}_j = \sum_{i \in \mathcal{P}} \lambda_{ij} \boldsymbol{A}_i, \tag{11.1}$$

where $\boldsymbol{A}_p$ is the row of $\boldsymbol{A}$ corresponding to place p and $\lambda_{ij} \in \mathbb{Q}$.

Proof. For every $\boldsymbol{m} \in \mathcal{M}(\mathcal{PN}, \boldsymbol{m}_0)$, $\exists \sigma$ such that $\boldsymbol{m}_0 | \sigma > \boldsymbol{m}$, which implies $\boldsymbol{m} = \boldsymbol{m}_0 + \boldsymbol{A}\bar{\boldsymbol{\sigma}}$. From (11.1), for all $j \in P$:

$$\begin{aligned} m(j) &= m_0(j) + \boldsymbol{A}_j \bar{\boldsymbol{\sigma}} = m_0(j) + \sum_{i \in \mathcal{P}} \lambda_{ij} \boldsymbol{A}_i \bar{\boldsymbol{\sigma}} \\ &= m_0(j) + \sum_{i \in \mathcal{P}} \lambda_{ij} (m(i) - m_0(i)). \end{aligned} \tag{11.2}$$

Conversely, assume $\exists j \in P \setminus \mathcal{P}$ such that (11.1) does not hold. Then, there exists a vector $\boldsymbol{v}$ which is perpendicular to the rows $\boldsymbol{A}_i, i \in \mathcal{P}$, but not to $\boldsymbol{A}_j$, i.e., $\exists \boldsymbol{v} \in \mathbb{Q}^N$ with $\boldsymbol{A}_i \boldsymbol{v} = 0, \forall i \in \mathcal{P}$, and $\boldsymbol{A}_j \boldsymbol{v} = 1$. For such $\boldsymbol{v}$, consider the firing sequence σ with count vector $\bar{\boldsymbol{\sigma}} = c\boldsymbol{v} + \sum_{i=1}^{h} \alpha_i \boldsymbol{x}^i$, with $c \in \mathbb{Z}/\{0\}$, $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^h$ the T-invariants of the net and $\alpha_i \in \mathbb{N}$. Consider the initial marking $\boldsymbol{m}_0^\sigma$ from which firing σ yields $\boldsymbol{m}^\sigma$. We have $m^\sigma(i) = m_0^\sigma(i) + \boldsymbol{A}_i \bar{\boldsymbol{\sigma}} = m_0^\sigma(i)$ for all $i \in \mathcal{P}$, while the markings $\boldsymbol{m}^\sigma$ and $\boldsymbol{m}_0^\sigma$ are different because

$$m^\sigma(j) = m_0^\sigma(j) + \boldsymbol{A}_j \bar{\boldsymbol{\sigma}} = m_0^\sigma(j) + \boldsymbol{A}_j \left(cx + \sum_{i=1}^{h} \alpha_i \boldsymbol{x}^i \right) = m_0^\sigma(j) + c.$$

Therefore, if (11.1) does not hold, $\mathcal{P}$ cannot be a sufficient place set. □

The sufficient place set of a Petri net (and the corresponding surplus place set) is in general not unique. Sufficient places sets, and thus surplus place sets, can be characterized from the P-invariants, since the linear relations between the rows of $\boldsymbol{A}$ are described by its P-invariants. This is also intuitive, because P-invariants characterize a constant weighted marking over a subset of places (see Definition 9.15).

Lemma 11.3. Consider a structurally live and structurally bounded Petri net. Let the set of its minimal support P-invariants be $\{\boldsymbol{y}^1, \ldots, \boldsymbol{y}^p\}$ and choose a place set $\mathcal{P} \subseteq P$. Whether $\mathcal{P}$ is a surplus place set can be characterized as follows:

Step 1. Obtain a basis $\{\bar{\boldsymbol{y}}^1,\ldots,\bar{\boldsymbol{y}}^r\}$ composed of elements from $\{\boldsymbol{y}^1,\ldots,\boldsymbol{y}^p\}$. Define matrix $\boldsymbol{Y}$ consisting of the rows $\{\bar{\boldsymbol{y}}^1,\ldots,\bar{\boldsymbol{y}}^r\}$.

Step 2. Order the columns of $\boldsymbol{Y}$ such that the columns according to places $p \in \mathcal{P}$ are in front. Denote the obtained matrix by $\widetilde{\boldsymbol{Y}}$.

Step 3. Apply Gauss-Jordan elimination on matrix $\widetilde{\boldsymbol{Y}}$ to obtain its reduced row echelon form $\mathrm{rref}(\widetilde{\boldsymbol{Y}})$.

Step 4. $\mathcal{P}$ is a surplus place set if and only if $\mathrm{rref}(\widetilde{\boldsymbol{Y}})$ contains leading ones in columns $1,\ldots,|\mathcal{P}|$.

Now, if $\mathcal{P}$ is a surplus place set, the marking of the places $j \in \mathcal{P}$ is expressed by the marking of the places $\mathcal{P}^{suf} = P\backslash\mathcal{P}$ as follows:

$$m(j) = m_0(j) - \sum_{i\in\mathcal{P}^{suf}} \mathrm{rref}(\widetilde{Y})_{ji}(m(i) - m_0(i)). \tag{11.3}$$

Proof. Let $\widetilde{\boldsymbol{A}}$ be the permutation of $\boldsymbol{A}$ corresponding to the permutation applied to obtain $\widetilde{\boldsymbol{Y}}$. Since $\boldsymbol{Y}\boldsymbol{A} = 0$, also $\widetilde{\boldsymbol{Y}}\widetilde{\boldsymbol{A}} = 0$ and $\mathrm{rref}(\widetilde{\boldsymbol{Y}})\widetilde{\boldsymbol{A}} = 0$. If $\mathrm{rref}(\widetilde{\boldsymbol{Y}})$ has leading ones in the first $|\mathcal{P}|$ columns, setting $\lambda_{ij} = -\mathrm{rref}(\widetilde{\boldsymbol{Y}})_{ji}$ in (11.1) implies by Lemma 6.2 that $\mathcal{P}$ is a surplus place set. In addition, (11.3) follows from (11.2).

Conversely, if $\mathcal{P}$ is a surplus place set, from (11.1) we can find a $\boldsymbol{w}_i \in \mathbb{Q}^N$ for every $i \in \mathcal{P}$ such that $\boldsymbol{w}_i\widetilde{\boldsymbol{A}} = 0$ by taking $w_i(i) = 1$, $w_i(p) = 0$ for all $p \in \mathcal{P}\backslash\{i\}$, and $w_i(p) = \lambda_{ij}$ for all $p \in P\backslash\mathcal{P}$. From Result 9.27 follows that each such $\boldsymbol{w}_i$ is a linear combination of minimal support P-invariants. This implies $\boldsymbol{w}_i \in \mathrm{rowspan}(\boldsymbol{Y}) = \mathrm{rowspan}(\widetilde{\boldsymbol{Y}})$ and thus $\boldsymbol{w}_i \in \mathrm{rowspan}(\mathrm{rref}(\widetilde{\boldsymbol{Y}}))$. Now assume that $\mathrm{rref}(\widetilde{\boldsymbol{Y}})$ does not have leading ones in the first $|\mathcal{P}|$ columns. Let j be the first column without a leading one and $\mathrm{rref}(\widetilde{\boldsymbol{Y}})_j$ the j-th row of $\mathrm{rref}(\widetilde{\boldsymbol{Y}})$. By showing that the equation

$$\boldsymbol{w}_j = \sum_{i=1}^{r} \alpha_i\, \mathrm{rref}(\widetilde{\boldsymbol{Y}})_j \tag{11.4}$$

has no solution, we obtain the contradiction $\boldsymbol{w}_j \notin \mathrm{rowspan}(\mathrm{rref}(\widetilde{\boldsymbol{Y}}))$, from which we conclude that $\mathrm{rref}(\widetilde{\boldsymbol{Y}})$ must have leading ones in the first $|\mathcal{P}|$ columns. $w_j(i) = 0$ for $i < j$ implies $\alpha_i = 0$ which reduces (11.4) to $\boldsymbol{w}_j = \sum_{i=j}^{r} \alpha_i \mathrm{rref}(\widetilde{\boldsymbol{Y}})_j$. Since $w_j(j) = 1$ the latter equation has no solution, because otherwise column j is a pivot column during the Gauss Jordan elimination, which would have resulted in a leading one in column j. □

Remark 11.4. Lemma 11.3 provides a test to check for a given candidate place set whether or not it is a surplus place set, since the columns of $\boldsymbol{Y}$ are pre-ordered. This test be used in the decomposition algorithm that we present at the end of this section. Observe that by starting from $\boldsymbol{Y}$ and applying Gauss-Jordan elimination while allowing swapping of columns, it is also possible to trace surplus place sets.

The minimal number of places a sufficient place set was already expressed (and defined as the *dimension of the marking process*) by Florin and Natkin [216]. From

each additional linearly independent P-invariant an additional surplus place can be selected. The number of linearly independent minimal support P-invariants is equal to $dim(Ker(\boldsymbol{A}^T))$. Recall that this number can be smaller than the number of minimal support P-invariants (see Remark 9.28).

Lemma 11.5 ([216])**.** For each sufficient place set $\mathcal{P}^{suf}$:

$$\left|\mathcal{P}^{suf}\right| \geq N - dim(Ker(\boldsymbol{A}^T)).$$

Remark 11.6. Note that the minimal number of places in a sufficient place set $\min\{\left|\mathcal{P}^{suf}\right|\}$ is directly connected to the notion of deficiency (discussed in Section 10.3.3): $\delta = |\mathcal{R}(\mathcal{T})| - \ell - Rank(\boldsymbol{A}) = |\mathcal{R}(\mathcal{T})| - \ell - (N - dim(Ker(\boldsymbol{A}^T))) = |\mathcal{R}(\mathcal{T})| - \ell - \min\{\left|\mathcal{P}^{suf}\right|\}$.

In Theorem 10.19, there may be solutions to the matrix equation $\boldsymbol{z}\boldsymbol{A} = \boldsymbol{C}$ with $z_p = 0$ for some places p. Such a place has $f_p = 1$ and no term involving place p appears in the product form (10.12). The following lemma shows that such places are uniquely related to places contained in a surplus place set.

Lemma 11.7. Assume a solution to the matrix equation $\boldsymbol{z}\boldsymbol{A} = \boldsymbol{C}$ exists. If $\mathcal{P}'$ is a surplus place set, then there exists a solution to $\boldsymbol{z}\boldsymbol{A} = \boldsymbol{C}$, where $z_p = 0$, for all $p \in \mathcal{P}'$ ($\mathcal{P}' \subseteq P$).

Proof. Consider a surplus set $\mathcal{P}'$. By Lemma 11.2, the row vectors $\boldsymbol{A}_j$ of $\boldsymbol{A}$ corresponding to the places $j \in \mathcal{P}'$ can be written as linear combination of the rows $\boldsymbol{A}_i, i \in P \backslash \mathcal{P}'$. Therefore, under the assumption that a solution $\boldsymbol{z}$ to $\boldsymbol{z}\boldsymbol{A} = \boldsymbol{C}$ exists, there exists a solution where $z_p = 0, \forall p \in \mathcal{P}'$. □

Firing of transitions of T-invariants of different common input bag classes interacts and conflicts in the places that are shared among the common input bag classes. Focusing on such places will enable us to formulate decomposition results. Therefore, we formally define *conflict places* and the set of all conflict places among all common input bag classes.

Definition 11.8 (**Conflict place - Conflict place set**)**.** Let CI^i and CI^j be two common input bag classes such that $i \neq j$. Let p be a place that is an element of both CI^i and CI^j, i.e., $p \in (\mathcal{P}(CI^i) \cap \mathcal{P}(CI^j))$. Then p is called a *conflict place* of CI^i and CI^j. The *conflict place set* is the subset $\mathcal{P}^{con} \subseteq P$, of places that are a conflict place between any two common input bag classes:

$$\mathcal{P}^{con} = \left\{p \in P \mid \exists i, j, i \neq j \text{ with } p \in \left(\mathcal{P}(CI^i) \cap \mathcal{P}(CI^j)\right)\right\}.$$

11.3 Decomposition

The decomposition result will be obtained by removing conflict places. Therefore, before stating the decomposition result, the following lemma is presented.

Lemma 11.9. If in an $\mathcal{S}\Pi$-net $\mathcal{SPN}$ the places and all arcs incident to all the places $p \in \mathcal{P} \subset P$ can be removed so that no complete input bag is removed, then the remaining net is an $\mathcal{S}\Pi$-net, possibly consisting of several separated components.

Proof. Remove from $\mathcal{SPN}$ a place $p' \in \mathcal{P}$ and the arcs incident to this place. There is no transition for which has $\boldsymbol{I}_p(t) = 0$ for all $p \in P \backslash p'$, since by removing all places $p \in \mathcal{P}$ no complete input bag is removed. Denote the remaining net by $\mathcal{SPN}'$. $\mathcal{SPN}'$ only differs from $\mathcal{SPN}$ in the transitions incident to place p'. We need to show that these transitions are still covered by minimal closed support T-invariants. Consider the set of minimal closed support T-invariants in $\mathcal{SPN}$ that visit place p', i.e., $\{\boldsymbol{x} \mid \exists t \in \|\boldsymbol{x}\| \text{ with } \boldsymbol{I}_{p'}(t) \geq 0 \vee \boldsymbol{O}_{p'}(t) \geq 0\}$. Now consider the consecutive transitions $t, t' \in \|\boldsymbol{x}\|$ for which $\boldsymbol{O}(t) = \boldsymbol{I}(t')$ and $\boldsymbol{O}_{p'}(t) \geq 0$ in the original net $\mathcal{SPN}$. In net $\mathcal{SPN}'$, $\boldsymbol{O}(t) = \boldsymbol{I}(t')$ still holds, since both in $\boldsymbol{O}(t)$ and $\boldsymbol{I}(t')$ place p' is removed. Therefore, each minimal closed support T-invariant $\boldsymbol{x}$ in $\mathcal{SPN}$ is still a minimal closed support T-invariant in $\mathcal{SPN}'$. Since it may be that for two minimal closed support T-invariants $\boldsymbol{x}^1, \boldsymbol{x}^2$ that both visit place p', place p' is the only conflict place of $CI(\boldsymbol{x}^1)$ and $CI(\boldsymbol{x}^2)$, i.e., $CI(\boldsymbol{x}^1) \cap CI(\boldsymbol{x}^2) = p'$, $\mathcal{SPN}'$ may consist of two separate $\mathcal{S}\Pi$-nets. The proof is completed by repeating this argument until all places $p \in \mathcal{P}$ are removed. □

If there is no transition for which the complete input bag is contained in the intersection of the surplus place set and conflict place set, i.e.,

Theorem 11.10. Consider a product form $\mathcal{SPN}$ and a surplus place set $\mathcal{P}^{sur}$ with corresponding sufficient place set $\mathcal{P}^{suf}$. If $\nexists t \in T$ for which $\{p \in P \mid I_p(t) > 0\} \subseteq \mathcal{P}^{int} = \{p \in P \mid p \in (\mathcal{P}^{con} \cap \mathcal{P}^{sur})\}$, then

- removing all places $p \in \mathcal{P}^{int}$ and all arcs incident to the places $p \in \mathcal{P}^{int}$ yields s product form $\mathcal{S}\Pi$-nets: $\mathcal{SPN}^1, \ldots, \mathcal{SPN}^s$; each $\mathcal{SPN}^i$ corresponding of one or more connected common input bag classes.
- the equilibrium distribution π of $\mathcal{SPN}$ is a product over the invariant measures of the subnets:

$$\pi(\boldsymbol{m}) = B \prod_{i=1}^{s} \pi_y^{\mathcal{SPN}^i}(\boldsymbol{m}^i), \quad \boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0),$$

where $\boldsymbol{m}^i$ is the submarking in places that belong to subnet $\mathcal{SPN}^i$, $\pi_y^{\mathcal{SPN}^i}(\boldsymbol{m}^i)$ is the invariant measure of subnet $\mathcal{SPN}^i$ with

$$\pi_y^{\mathcal{SPN}^i}(\boldsymbol{m}^i) = \prod_{\{p \in \cap_{j=1}^{J^i} \mathcal{P}(CI^j) \backslash \mathcal{P}^{con}\}} f_p^{m_p}, \tag{11.5}$$

where $CI^j, j = 1, \ldots, J^i$, denote the J^i common input bag classes contained in subnet $\mathcal{SPN}^i$, and B is a normalizing constant with $B^{-1} = \sum_{\boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)} \pi_y(\boldsymbol{m})$.

Proof. When the places $p \in \mathcal{P}^{int}$ and all arcs connected to these places are removed from $\mathcal{SPN}$, by Lemma 11.9, $\mathcal{SPN}$ falls apart in subnets $\mathcal{SPN}^1, \ldots, \mathcal{SPN}^s$ that are again $\mathcal{S}\Pi$-nets. Since in general not all conflict places are contained in $\mathcal{P}^{int}$, common input bag classes that share a conflict place that is not contained in $\mathcal{P}$ are contained in the same subnet $\mathcal{SPN}^i$.

For the second part, by Lemma 11.7, for $\mathcal{SPN}$ there exists a solution to $\boldsymbol{z}\boldsymbol{A} = \boldsymbol{C}$, in which $z_p = 0, \forall p \in \mathcal{P}^{con}$. The product form stationary distribution (10.12) can thus be rewritten as

$$\pi_y(\boldsymbol{m}) = \prod_{i=1}^{s} \left\{ \prod_{\{p \in \cap_{j=1}^{J^i} \mathcal{P}(CI^j) \setminus \mathcal{P}^{con}\}} f_p^{m_p} \right\}.$$

We are left to show that the f_p values of the subnets are equal to those of the original net. This can be seen as follows. Introduce matrix $\boldsymbol{A}'$, which is the modified incidence matrix $\boldsymbol{A}$ so that the rows corresponding to the places of the conflict place set are set to zero, i.e., $\boldsymbol{a}_p = 0$ for all $p \in \mathcal{P}^{con}$. Then we have $\boldsymbol{z}\boldsymbol{A} = \boldsymbol{z}\boldsymbol{A}'$. The system of equations $\boldsymbol{z}\boldsymbol{A}' = \boldsymbol{C}$ can be permuted such that that the conflict places are grouped and the places of each $\mathcal{SPN}^i$ class are grouped:

$$\widetilde{\boldsymbol{z}}\widetilde{\boldsymbol{A}'} = \widetilde{\boldsymbol{z}} \begin{pmatrix} \boldsymbol{A}^1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{A}^2 & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ \vdots & \cdots & 0 & \boldsymbol{A}^s \\ 0 & \cdots & \cdots & 0 \end{pmatrix} = \widetilde{\boldsymbol{C}} = \begin{pmatrix} \boldsymbol{C}^1 & \cdots & \boldsymbol{C}^s \end{pmatrix}.$$

The proof is concluded by observing that the matrices $\boldsymbol{A}^i$ and vectors $\boldsymbol{C}^i, i = 1, \ldots, s$ correspond exactly to the incidence matrices and the $\boldsymbol{C}$-vectors of the subnets $\mathcal{SPN}^1$, $\ldots, \mathcal{SPN}^s$. □

Below, we will argue that Theorem 11.10 is a generalization of Frosch and Natarajan [223]. This will be further illustrated along several example Petri nets in Section 11.4. Let us first provide the formal definition of a $\mathcal{CS}$ and provide the theorem of Frosch and Natarajan.

Definition 11.11 (Closed Synchronized Systems of Stochastic Sequential Processes ($\mathcal{CS}$))**.** A structurally bounded stochastic Petri net $\mathcal{SPN} = (P_1 \cup \ldots \cup P_m \cup \mathcal{B}, T_1 \cup \ldots \cup T_m, I, O, Q)$ is a closed synchronized system of stochastic sequential processes if and only if:

1. $\forall i, j \in \{1, \ldots, m\}$ such that $i \neq j : P_i \cap P_j = \emptyset, T_i \cap T_j = \emptyset, P_i \cap \mathcal{B} = \emptyset$,
2. $\forall i \in \{1, \ldots, m\} : \mathcal{M}_i = (P_i, T_i, I|_i, O|_i, Q|_i)$ are cyclic state machines (where $I|_i, O|_i, Q|_i$ are the restrictions of I, O and Q to P_i and T_i).

Theorem 11.12 ([223])**.** Let $(\mathcal{SPN}, \boldsymbol{m}_0)$ be a live and marked $\mathcal{CS}$. Consider the following assumption:

$\mathcal{A}$: Let $\boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)$ and t_0 a transition in state machine $\mathcal{M}_i$, which is enabled in $\boldsymbol{m}$. Further, let $\boldsymbol{x}$ be a minimal support T-invariant of $\mathcal{M}_i$ such that $t_0 \in \|\boldsymbol{x}\|$. Then the sequential transition sequence $\sigma = (t_0, t_1, \ldots, t_n)$ in $\mathcal{M}_i$ corresponding to $\boldsymbol{x}$ has to be a firing sequence in $\boldsymbol{m}$, i.e. $\boldsymbol{m}[\sigma > \boldsymbol{m}' \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)$.

Let $(\mathcal{SPN}, \boldsymbol{m}_0)$ satisfy $\mathcal{A}$. Then the equilibrium distribution π of $(\mathcal{SPN}, \boldsymbol{m}_0)$ is given by

$$\pi(\boldsymbol{m}) = B \prod_{i=1}^{m} \pi_y^{\mathcal{SPN}^i}(\boldsymbol{m}^i), \quad \boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0),$$

where B is a normalizing constant and $\pi_y^{\mathcal{SPN}^i}(\boldsymbol{m}^i)$ is the invariant measure of state machine i.

First, a $\mathcal{CS}$ is obtained by starting from separate state machines and linking these by buffer places, so that the buffer places are defined beforehand. Therefore, Theorem 11.12 can be interpreted as a composition result rather than a decomposition result. In addition, note that it not a structural decomposition result, but a behavioral one.

Second, Assumption $\mathcal{A}$ ensures that the connection of the state machines is such that the state machines are synchronized by the buffer places in a way that the transitions of the state machines are expanded with arcs to the buffer places so that only minimal closed support T-invariants are formed from the T-invariants of the state machines. As a consequence, a $\mathcal{CS}$ that satisfies assumption $\mathcal{A}$ is an $\mathcal{S}\Pi^2$-net. Theorem 11.10 is not restricted to $\mathcal{S}\Pi^2$-nets.

To conclude, we present an algorithm that exploits Lemma 11.3 and Theorem 11.10 to find all possible decompositions of an $\mathcal{S}\Pi$-net. Observe that decomposition according to Theorem 11.10 is realized by identifying places that are both conflict places and surplus places. In the algorithm below we exploit this property, by generating surplus place sets that are contained in the conflict place set. Also observe that removing the places in $\mathcal{P}^{int}$ in Theorem 11.10 either removes a complete input bag or implies a decomposition. In addition, since a sufficient place set is in general not unique, the decomposition according to Theorem 11.10 is not unique. Each surplus place set that provides a decomposition, provides a specific decomposition. However, different surplus place sets may lead to the same decomposition if they have an identical intersection with the conflict place set.

Algorithm 11.13 (Generating all decompositions)**.**

Step 1. *Consider a product form $\mathcal{SPN}$. Execute the following initialization steps:*

(a) *Determine from the set of common input bag classes $\mathcal{C} = \{CI^1, \ldots CI^\ell\}$, the set of conflict places:*

$$\mathcal{P}^{con} = \left\{ p \in P \mid p \in \left(\mathcal{P}(CI^i) \cap \mathcal{P}(CI^j) \right), \forall i, j \text{ such that } i \neq j \right\}.$$

(b) *Obtain the powerset* $\mathcal{P}^{con}_{all} = Power(\mathcal{P}^{con})$ *of the set* $\mathcal{P}^{con}$*. Remove from* $\mathcal{P}^{con}_{all}$ *all sets that contain a complete input bag.*

(c) *Define the set of surplus place sets that provide a decomposition* $\mathcal{P}^{dec}_{all}$ *and set* $\mathcal{P}^{dec}_{all} = \emptyset$*.*

Step 2. *Take an element* $\mathcal{P} \in \mathcal{P}^{con}_{all}$ *and apply the procedure from Lemma 11.3 to check whether* $\mathcal{P}$ *is a surplus place set. If yes, go to step 3, else go to step 4.*

Step 3. $\mathcal{P}^{con}_{all} := \mathcal{P}^{con}_{all} \backslash Power(\mathcal{P})$ *and* $\mathcal{P}^{dec}_{all} := \mathcal{P}^{dec}_{all} \cup Power(\mathcal{P})$*. Go to step 5.*

Step 4. *Remove* $\mathcal{P}$ *and all its supersets from* $\mathcal{P}^{con}_{all}$*, i.e.,*

$$\mathcal{P}^{con}_{all} := \mathcal{P}^{con}_{all} \backslash \{\mathcal{P}' | \mathcal{P}' \in \mathcal{P}^{con}_{all}, \mathcal{P} \subseteq \mathcal{P}'\}.$$

Step 5. *If* $\mathcal{P}^{con}_{all} \neq \emptyset$ *go back to step 2, else go to step 6.*

Step 6. *For each surplus place set* $\mathcal{P} \in \mathcal{P}^{dec}_{all}$*, solving* $z\boldsymbol{A} = \boldsymbol{C}$ *with* $z_p = 0$ *for* $p \in \mathcal{P}$*, yields a unique decomposition of the equilibrium distribution of* $\mathcal{SPN}$*:*

$$\pi(\boldsymbol{m}) = B \prod_{i=1}^{s} \pi_y^{\mathcal{SPN}^i}(\boldsymbol{m}^i), \text{ with } \pi_y^{\mathcal{SPN}^i}(\boldsymbol{m}^i) \text{ given in (11.5).}$$

11.4 Examples

To illustrate Theorem 11.10, we present three examples. First, in Example 11.14, all conflict places can be removed, which implies a decomposition that separates all common input bag classes. Second, Example 11.15 presents a net with a decomposition where several common input bag classes stay connected, because it is not allowed that a complete input bag is contained in $\mathcal{P}^{int}$. Otherwise, at least one of the minimal closed support T-invariants would be removed. Example 11.15 also provides an illustration of the algorithm to obtain all possible decompositions. Both in Example 11.14 and 11.15 buffer places can be identified so that they fall within the $\mathcal{CS}$ class of (de)composable $\mathcal{SPN}$s according to Frosch and Natarajan [223]. Example 11.16 shows that Theorem 11.10 is a generalization of Theorem 11.12, by presenting a decomposable $\mathcal{S}\Pi$-net which is not a $\mathcal{CS}$.

Example 11.14 (Complete decomposition in common input bag classes)**.** Consider the Petri net depicted in Figure 11.1. From the incidence matrix

$$A = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ -1 & 1 & -1 & 1 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 \end{pmatrix},$$

we obtain that this net has two T-invariants $\boldsymbol{x}^1 = (1100)$ and $\boldsymbol{x}^2 = (0011)$ and three minimal support P-invariants $\boldsymbol{y}^1 = (11000)$, $\boldsymbol{y}^2 = (00011)$ and $\boldsymbol{y}^3 = (01101)$, which are linearly independent. The number of places in a sufficient place set is thus

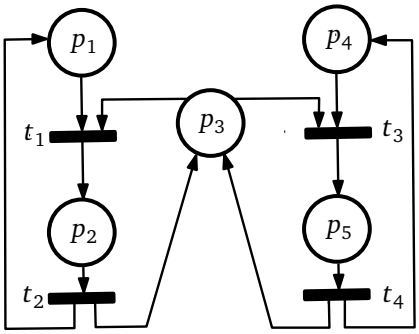


Figure 11.1: An $\mathcal{SPN}$ decomposing into all its common input bag classes.

$N - 3 = 2$. The two minimal support T-invariants both have a closed support, so that it is an $\mathcal{S}\Pi^2$-net, and $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$ are not in common input bag relation, so that we have common input bag classes $CI(\boldsymbol{x}^1)$ and $CI(\boldsymbol{x}^2)$, with one conflict place p_3.

Consider the sufficient place set $\mathcal{P}^{suf} = \{p_1, p_4\}$, with corresponding surplus place set $\mathcal{P}^{sur} = \{p_2, p_3, p_5\}$. Then, the conditions of Theorem 11.10 are satisfied, and by removing place p_3 the net decomposes into two subnets: $\mathcal{SPN}^1$ related to $CI(\boldsymbol{x}^1)$ and $\mathcal{SPN}^2$ related to $CI(\boldsymbol{x}^2)$, with invariant measures

$$\pi_y^{\mathcal{SPN}^1}(\boldsymbol{m}^1) = \left(\frac{\mu_2}{\mu_1}\right)^{m_1} \quad \text{and} \quad \pi_y^{\mathcal{SPN}^2}(\boldsymbol{m}^2) = \left(\frac{\mu_4}{\mu_3}\right)^{m_4}.$$

The equilibrium distribution of $\mathcal{SPN}$ is

$$\pi(\boldsymbol{m}) = B\pi_y^{\mathcal{SPN}^1}(\boldsymbol{m}^1)\pi_y^{\mathcal{SPN}^2}(\boldsymbol{m}^2), \qquad \boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0).$$

This example is an illustration of a special case of Theorem 11.10. Recall that $\{CI^1, \ldots, CI^\ell\}$ is the set of common input bag classes of a certain $\mathcal{S}\Pi$-net $\mathcal{SPN}$. When for $\mathcal{SPN}$ there exists a surplus place set $\mathcal{P}^{sur}$ and corresponding sufficient place set $\mathcal{P}^{suf}$, such that $\mathcal{P}^{con} \subseteq \mathcal{P}^{sur}$ and $\nexists t \in T$ for which $\{p \in P \mid \boldsymbol{I}_p(t) > 0\} \subseteq \mathcal{P}^{con}$, $\mathcal{SPN}$ decomposes in ℓ subnets $\mathcal{SPN}^1, \ldots, \mathcal{SPN}^\ell$ with each of these subnets corresponding to one common input bag class CI^i. □

Example 11.15 (Decomposition in connected common input bag classes)**.** Consider the Petri net depicted in Figure 11.2. From the incidence matrix

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix},$$

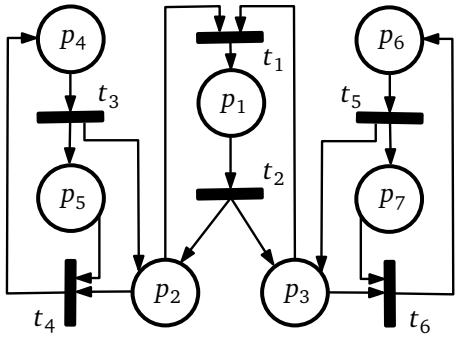


Figure 11.2: An $\mathcal{SPN}$ that decomposes into two components.

we obtain that this net has three minimal support T-invariants $\boldsymbol{x}^1 = (110000)$, $\boldsymbol{x}^2 = (001100)$ and $\boldsymbol{x}^3 = (000011)$ and four minimal support P-invariants $\boldsymbol{y}^1 = (0001100)$, $\boldsymbol{y}^2 = (0000011)$, $\boldsymbol{y}^3 = (1101000)$, and $\boldsymbol{y}^4 = (1010010)$, which are linearly independent. The number of places in a minimal sufficient place set is thus $N - 4 = 3$. The three minimal support T-invariants all have a closed support, so that it is an $\mathcal{S}\Pi^2$-net, and $\boldsymbol{x}^1$, $\boldsymbol{x}^2$ and $\boldsymbol{x}^3$ are not in common input bag relation, so that we have common input bag classes $CI(\boldsymbol{x}^1) = \{\boldsymbol{x}^1\}$, $CI(\boldsymbol{x}^2) = \{\boldsymbol{x}^2\}$ and $CI(\boldsymbol{x}^3) = \{\boldsymbol{x}^3\}$. The conflict place set is $\mathcal{P}^{con} = \{p_2, p_3\}$. The complete input bag of transition t_1 is contained in the conflict set, so that not all conflict places can be removed.

However, the connection of common input bag class $CI(\boldsymbol{x}^3)$ with the rest of the network is such that it can be decomposed from the network. Note that for a given sufficient place set $\mathcal{P}^{suf}$ and corresponding surplus place set $\mathcal{P}^{sur}$, $\mathcal{P}' = (\mathcal{P}^{suf} \cup p)$ with $p \in \mathcal{P}^{sur}$ is also a sufficient place set. Therefore, choose $\mathcal{P}^{sur} = \{p_3, p_5, p_7\}$, so that $(\mathcal{P}^{sur} \cap \mathcal{P}^{con}) = \{p_3\}$. By Theorem 11.10 the network decomposes into $\mathcal{SPN}^1 = \{CI(\boldsymbol{x}^1), CI(\boldsymbol{x}^2)\}$ and $\mathcal{SPN}^2 = \{CI(\boldsymbol{x}^3)\}$, with invariant measures

$$\pi_y^{\mathcal{SPN}^1}(\boldsymbol{m}^1) = \left(\frac{\mu_1}{\mu_2}\right)^{m_1} \left(\frac{\mu_4}{\mu_3}\right)^{m_4} \quad \text{and} \quad \pi_y^{\mathcal{SPN}^2}(\boldsymbol{m}^2) = \left(\frac{\mu_6}{\mu_5}\right)^{m_6}.$$

The equilibrium distribution of $\mathcal{SPN}$ is

$$\pi(\boldsymbol{m}) = B\pi_y^{\mathcal{SPN}^1}(\boldsymbol{m}^1)\pi_y^{\mathcal{SPN}^2}(\boldsymbol{m}^2) \qquad , \boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0).$$

Observe that a decomposition in $\mathcal{SPN}^1 = \{CI(\boldsymbol{x}^1), CI(\boldsymbol{x}^3)\}$, $\mathcal{SPN}^2 = \{CI(\boldsymbol{x}^2)\}$ would be possible too. Algorithm 11.13 allows us to find the different decomposition possibilities. To illustrate its application we execute the algorithm to this simple but insightful example.

Step 1. The conflict place set is $\mathcal{P}^{con} = \{p_2, p_3\}$. Therefore, the candidate decomposition place sets are $\{p_2\}, \{p_3\}$ and $\{p_2, p_3\}$, from which $\{p_2, p_3\}$ is removed as it contains a complete input bag. Thus, $\mathcal{P}^{con}_{all} = \{\{p_2\}, \{p_3\}\}$.

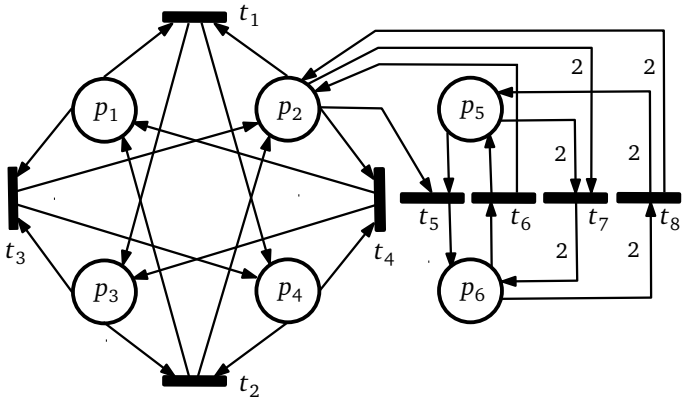


Figure 11.3: A decomposable $\mathcal{SPN}$ which is neither a $\mathcal{CS}$ nor an $\mathcal{S}\Pi^2$-net.

Step 2-5. Both $\{p_2\}$ and $\{p_3\}$ are surplus place sets. As a consequence, there are two options to decompose the $\mathcal{SPN}$: $\mathcal{P}_{all}^{dec} = \{\{p_2\}, \{p_3\}\}$.

Step 6. The two possible decompositions both divide the $\mathcal{SPN}$ in two subnetworks such that

$$\pi(\boldsymbol{m}) = B\pi_y^{\mathcal{SPN}^1}(\boldsymbol{m}^1)\pi_y^{\mathcal{SPN}^2}(\boldsymbol{m}^2), \qquad \boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0),$$

where for the first surplus place set $\{p_2\}$ the two subnetworks are $\mathcal{SPN}^1 = \{CI(\boldsymbol{x}^1, \boldsymbol{x}^3)\}$ and $\mathcal{SPN}^2 = \{CI(\boldsymbol{x}^2)\}$ and for the second surplus place set $\{p_3\}$ these are $\mathcal{SPN}^1 = \{CI(\boldsymbol{x}^1, \boldsymbol{x}^2)\}$ and $\mathcal{SPN}^2 = \{CI(\boldsymbol{x}^3)\}$. □

Example 11.16 (Non-$\mathcal{CS}$, Non-$\mathcal{S}\Pi^2$)**.** Both Petri nets from Example 11.14 and Example 11.15 can be regarded as $\mathcal{CS}$s, when the buffer places in $\mathcal{B}$ are respectively chosen as $\{p_3\}$ and $\{p_3, p_5\}$. Now, consider the stochastic Petri net $\mathcal{SPN}$ depicted in Figure 11.3. This is an example of an $\mathcal{S}\Pi$-net, which is neither a $\mathcal{CS}$, the class of decomposable $\mathcal{SPN}$s defined by Frosch and Natarajan [223], nor an $\mathcal{S}\Pi^2$-net. From the incidence matrix

$$A = \begin{pmatrix} -1 & 1 & -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 1 & -1 & -1 & 1 & -2 & 2 \\ 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & -2 & 2 \\ 0 & 0 & 0 & 0 & 1 & -1 & 2 & -2 \end{pmatrix},$$

we obtain that this net has six minimal support T-invariants $\boldsymbol{x}^1 = (11000000)$, $\boldsymbol{x}^2 = (00110000)$, $\boldsymbol{x}^3 = (00001100)$, $\boldsymbol{x}^4 = (00000011)$, $\boldsymbol{x}^5 = (00000210)$ and $\boldsymbol{x}^6 = (00002001)$, of which $\boldsymbol{x}^1, \boldsymbol{x}^2, \boldsymbol{x}^3$ and $\boldsymbol{x}^4$ have a closed support. It has three minimal support P-invariants $\boldsymbol{y}^1 = (100100)$, $\boldsymbol{y}^2 = (011001)$ and $\boldsymbol{y}^3 = (000011)$,

which are linearly independent. The number of places in a sufficient place set is thus $N - 3 = 3$.

The minimal closed support T-invariants $\boldsymbol{x}^1, \boldsymbol{x}^2, \boldsymbol{x}^3, \boldsymbol{x}^4$ are not in common input bag relation, so that we have common input bag classes $CI(\boldsymbol{x}^1), CI(\boldsymbol{x}^2), CI(\boldsymbol{x}^3)$ and $CI(\boldsymbol{x}^4)$ with conflict place set $\mathcal{P}^{con} = \{p_1, p_2, p_3, p_4, p_5, p_6\}$. Since $\mathcal{SPN}$ is not an $\mathcal{S}\Pi^2$-net, for product form an additional condition on the numerical values of the transition rates is imposed, which is $(\mu_5/\mu_6)^2 = \mu_7/\mu_8$.

$CI(\boldsymbol{x}^1)$ and $CI(\boldsymbol{x}^2)$ cannot be disconnected according to Theorem 11.10, since it would require removal of a complete output bag. The same holds for $CI(\boldsymbol{x}^3)$ and $CI(\boldsymbol{x}^4)$. Therefore, consider the surplus place set $\mathcal{P}^{sur} = \{p_2\}$, with corresponding sufficient place set $\mathcal{P}^{suf} = \{p_1, p_3, p_4, p_5, p_6\}$. Then the conditions of Theorem 11.10 are satisfied, and by removing place p_2 the net decomposes in two subnets: $\mathcal{SPN}^1$ related to $CI(\boldsymbol{x}^1)$ and $CI(\boldsymbol{x}^2)$, and $\mathcal{SPN}^2$ related to $CI(\boldsymbol{x}^3)$ and $CI(\boldsymbol{x}^4)$, with invariant measures

$$\pi_y^{\mathcal{SPN}^1}(\boldsymbol{m}^1) = \left(\frac{\mu_1\mu_4}{\mu_2\mu_3}\right)^{\frac{1}{2}(m_1+m_3)} \left(\frac{\mu_1}{\mu_2}\right)^{m_4}, \text{and}$$

$$\pi_y^{\mathcal{SPN}^2}(\boldsymbol{m}^2) = \left(\frac{1}{\mu_5}\right)^{m_5} \left(\frac{1}{\mu_6}\right)^{m_6}.$$

The equilibrium distribution of $\mathcal{SPN}$ is

$$\pi(\boldsymbol{m}) = B\pi_y^{\mathcal{SPN}^1}(\boldsymbol{m}^1)\pi_y^{\mathcal{SPN}^2}(\boldsymbol{m}^2), \qquad \boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0).$$

To conclude, observe that an example of a decomposable $\mathcal{S}\Pi$-net which is not a $\mathcal{CS}$, but which is an $\mathcal{S}\Pi^2$-net, would be $\mathcal{SPN}$ without transitions t_7 and t_8. □

CHAPTER 12

Structural Decomposition via Bag Count Places

12.1 Introduction

Chapter 11 described a structural decomposition result for $\mathcal{S}\Pi$-nets formulated exclusively in terms of P- and T-invariants using so-called conflict places (places that are shared by different minimal closed support T-invariants) and surplus places (places that can be omitted in characterizing the marking of the Petri net). Using the P-invariants to assign conflict places as surplus places, an algorithmic procedure was formulated to decompose a product form stochastic Petri net into subnets. The subnets corresponded to one or more common input bag classes, the equivalence classes of T-invariants of the stochastic Petri nets that share an input bag.

In the present chapter, we take the results from Chapter 11 as starting point to formulate an additional decomposition result. We focus on the subclass of $\mathcal{S}\Pi$-nets that have a product form equilibrium distribution irrespective of the transition rates. These nets where algebraically characterized by Haddad et al. [270] as $\mathcal{S}\Pi^2$-nets (see Definition 10.22), and in Chapter 10 we showed that $\mathcal{S}\Pi^2$-nets are the nets in which each minimal support T-invariant is a closed support T-invariant. We will present a decomposition theorem by which all $\mathcal{S}\Pi^2$-nets can be separated in all their common input bag classes.

We build on the characterization of $\mathcal{S}\Pi^2$-nets provided by Haddad et al. [270], by establishing an interpretation of the vectors $\boldsymbol{a}_{\boldsymbol{r}}$ that can be calculated for each bag $\boldsymbol{r} \in \mathcal{R}(T)$ according to Definition 10.22. Starting from an arbitrary $\mathcal{S}\Pi^2$-net, and introducing 'bag count places', we introduce the Bag-Count-Place-Extended Petri net of an $\mathcal{S}\Pi^2$-net ($\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net). The Petri net that is formed by exclusively the bag count places consists of a set of state machine, one state machine per common input bag class. Along the concept of bag count places we show that the $\boldsymbol{a}_{\boldsymbol{r}}$-vectors provide the explicit relation between a marking difference $\boldsymbol{m} - \boldsymbol{m}'$ and the number of times each bag $\boldsymbol{r}$ is used in a firing sequence that is associated with this marking difference. This relation induces a one-to-one correspondence between the marking of the original places and the additionally constructed bag count places.

The one-to-one correspondence implies that the bag count places of a $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$ form a sufficient place set (see Definition 11.1), and thus that the equilibrium dis-

tribution of the bag count places provides an equilibrium distribution of the original places. In addition, by construction the bag count places a $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net are non-conflict places (see Definition 11.8). This enables us to apply Theorem 11.10 to the $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net. We obtain that the invariant measure of any $\mathcal{S}\Pi^2$-net factorizes in the invariant measures of the separate state machines that are associated with each of the common input bag classes.

The chapter is organized as follows. Section 12.2 defines the bag count places, introduces $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-nets, and discusses the interpretation of the $\boldsymbol{a}_r$-vectors. Next, Section 12.3 formulates the decomposition result, and Section 12.4 provides several examples.

12.2 Bag count places

This section introduces the Bag-Count-Place-Extended Petri net of a bounded $\mathcal{S}\Pi^2$-net. For every input/output bag of an $\mathcal{S}\Pi^2$-net a 'bag count' place is added to the original net. By connecting the bag count places to the existing transitions, the marking of these places will track the marking of the original places by counting the net number of times each bag $\boldsymbol{r} \in \mathcal{R}(T)$ is consumed and deposited. It will be shown that the $\boldsymbol{a}_r$-vectors from Definition 10.22 induce a one-to-one correspondence between the marking of the original places and the bag count places.

Definition 12.1 ($\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net)**.** Let $\mathcal{SPN} = (P, T, I, O, Q)$ be a structurally bounded $\mathcal{S}\Pi^2$-net. For each $\boldsymbol{r} \in \mathcal{R}(T)$, add bag count place $p^*_{\boldsymbol{r}}$ to P. The Bag-Count-Place-Extended $\mathcal{S}\Pi^2$-net ($\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net) of $\mathcal{SPN}$ is $\mathcal{SPN}^* = (\bar{P}, T, \bar{I}, \bar{O}, Q)$, where

- $\bar{P} = P \cup \mathcal{P}^*$, with $\mathcal{P}^* = \bigcup_{\boldsymbol{r} \in \mathcal{R}(T)} p^*_{\boldsymbol{r}}$,
- $\bar{I}, \bar{O} : \bar{P} \times T \to \mathbb{N}$ with

$$\bar{I}(p,t) = \begin{cases} I(p,t) & \text{, if } p \in P, \\ 1 & \text{, if } p = p^*_{\boldsymbol{r}}, \boldsymbol{r} = \boldsymbol{I}(t), \\ 0 & \text{, otherwise,} \end{cases}$$

and

$$\bar{O}(p,t) = \begin{cases} O(p,t) & \text{, if } p \in P. \\ 1 & \text{, if } p = p^*_{\boldsymbol{r}}, \boldsymbol{r} = \boldsymbol{O}(t), \\ 0 & \text{, otherwise.} \end{cases}$$

Note that the marking of a bag count place $p^*_{\boldsymbol{r}}$ changes if and only if a transition fires that either uses $\boldsymbol{r}$ as its input bag (in this case the marking of $p^*_{\boldsymbol{r}}$ decreases by one), or creates $\boldsymbol{r}$ as its output bag (in this case the marking of $p^*_{\boldsymbol{r}}$ increases by one). So the marking of $p^*_{\boldsymbol{r}}$ indicates the number of times bag $\boldsymbol{r}$ is created minus the number of times bag $\boldsymbol{r}$ is used. This insight is the starting point to obtain the marking

of the original places from the marking of the bag count places. To this end, first, in Lemma 12.2, we show that a $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net is an $\mathcal{S}\Pi^2$-net. Later, we will show in Lemma 12.5 that the initial marking on the bag count places can be chosen such that the marking of these places always remains positive, so that a $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net is an $\mathcal{SPN}$.

Definition is 12.1 a structural characterization. Lemmas 12.3 and 12.5 will show that for certain initial markings the behavior of a $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net is equivalent to its defining $\mathcal{S}\Pi^2$-net. Lemma 12.3 provides two conditions on the initial marking of the $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net which guarantee that a firing sequence σ can be fired in the original net if and only if σ can be fired in the $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net. Lemma 12.5 shows that for each structurally bounded $\mathcal{S}\Pi^2$-net, an initial marking satisfying the conditions of Lemma 12.3 can indeed be found. In Theorem 12.4 it is shown that there exists a one-to-one correspondence between the marking of the original places and the marking of the bag count places. Lemma 12.6 shows that a $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net is structurally bounded, a property that is a prerequisite for the decomposition result presented in Section 12.3. The decomposition result uses the result of Lemma 12.7 which gives the physical interpretation of $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-nets and therefore the $\boldsymbol{a}_r$-vectors in terms of state machines.

Lemma 12.2. The $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net $\mathcal{SPN}^*$ of an $\mathcal{S}\Pi^2$-net $\mathcal{SPN}$ is an $\mathcal{S}\Pi^2$-net.

Proof. Consider a minimal closed support T-invariant $\boldsymbol{x}$ of $\mathcal{SPN}$. For any transition $t \in \|\boldsymbol{x}\|$ there is a unique $t' \in \|\boldsymbol{x}\|$ such that $\boldsymbol{O}(t) = \boldsymbol{I}(t')$. By the construction of the $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net this yields $\boldsymbol{\alpha}(p^*, \boldsymbol{I}(t))\boldsymbol{x} = 0$, where $\boldsymbol{\alpha}(p^*, \boldsymbol{I}(t))$ denotes the row of the incidence matrix $\bar{\boldsymbol{A}}$ corresponding to place p_r^* with $\boldsymbol{r} = \boldsymbol{I}(t)$. Thus, $\boldsymbol{x}$ is also a T-invariant of $\mathcal{SPN}^*$. In addition, to see that $\boldsymbol{x}$ is a minimal closed support T-invariant of $\mathcal{SPN}^*$, observe that by construction if $\boldsymbol{I}(t) = \boldsymbol{O}(t')$ then $\bar{\boldsymbol{I}}(t) = \bar{\boldsymbol{O}}(t')$ also.

Next, every T-invariant of $\mathcal{SPN}^*$ is a T-invariant of $\mathcal{SPN}$, because the rows of $\bar{\boldsymbol{A}}$ for $p \in P$ are equal to the corresponding rows of $\boldsymbol{A}$, and thus, $\bar{\boldsymbol{A}}\boldsymbol{x} = 0 \Rightarrow \boldsymbol{A}\boldsymbol{x} = 0$. Thus, every minimal support T-invariant of $\mathcal{SPN}^*$ is a minimal closed support T-invariant.

Finally, since $\mathcal{SPN}$ and $\mathcal{SPN}^*$ have the same transition set T, it follows that in $\mathcal{SPN}^*$ every transition is covered by a minimal closed support T-invariant. □

Lemma 12.3. If the initial marking, $\bar{\boldsymbol{m}}_0$, of a $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net $\mathcal{SPN}^*$ corresponding to the marked $\mathcal{S}\Pi^2$-net $(\mathcal{SPN}, \boldsymbol{m}_0)$, is chosen such that $(\mathcal{SPN}^*, \bar{\boldsymbol{m}}_0)$ satisfies:

1. $\bar{m}_0(p) = m_0(p)$, for $p \in P$, and
2. for all $\bar{\boldsymbol{m}} \in \mathcal{M}(\mathcal{SPN}^*, \bar{m}_0)$, $\bar{m}(p) \geq 1$, for $p \in \mathcal{P}^*$,

then any firing sequence σ can be fired in $\mathcal{SPN}$ from $\boldsymbol{m}_0$ if and only if σ can be fired in $\mathcal{SPN}^*$ from $\bar{\boldsymbol{m}}_0$.

Proof. First, we show that every firing sequence σ that can be fired from $\boldsymbol{m}_0$ in $\mathcal{SPN}$ can also be fired from $\bar{\boldsymbol{m}}_0$ in $\mathcal{SPN}^*$. Since $\bar{I}(p,t) = I(p,t)$ and $\bar{O}(p,t) = O(p,t)$ for places $p \in P$, these places will never disable a transition that is enabled in $\mathcal{SPN}$.

Because $\bar{I}(p,t) \leq 1$ for $p \in \mathcal{P}^*$, condition 2. ensures that the same holds for these places.

Conversely, every firing sequence σ that can be fired from $\bar{\boldsymbol{m}}_0$ in $\mathcal{SPN}^*$ can be fired from $\boldsymbol{m}_0$ in $\mathcal{SPN}$, because $\bar{I}(p,t) = I(p,t)$ and $\bar{O}(p,t) = O(p,t)$ for places $p \in P$, and any transition $t \in T$ consumes and deposits the same number of tokens from the same places $p \in P$ in both nets. □

Theorem 12.4. Let $(\mathcal{SPN}^*, \bar{\boldsymbol{m}}_0)$ be a marked $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net corresponding to the marked $\mathcal{S}\Pi^2$-net $(\mathcal{SPN}, \boldsymbol{m}_0)$, and consider the markings $\boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)$ and $\bar{\boldsymbol{m}} \in \mathcal{M}(\mathcal{SPN}^*, \bar{\boldsymbol{m}}_0)$.

1. The marking of the places $p \in \mathcal{P}^*$ in the $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net can be expressed in terms of the marking of the places $p \in P$ as follows:

$$\bar{m}(p_r^*) = \bar{m}_0(p_r^*) + \boldsymbol{a}_r(\boldsymbol{m} - \boldsymbol{m}_0), \tag{12.1}$$

   where $\boldsymbol{a}_r$ is a vector as given in Definition 10.22.

2. The marking of the places $p \in P$ can be expressed in the marking of the places $p \in \mathcal{P}^*$ as follows:

$$\boldsymbol{m} = \boldsymbol{m}_0 + \sum_{r \in \mathcal{R}(T)} (\bar{m}(p_r^*) - \bar{m}_0(p_r^*))\boldsymbol{r}.$$

As a consequence, there is a unique relation between the marking $\boldsymbol{m}$ of $\mathcal{SPN}$ and $\bar{\boldsymbol{m}}$ of $\mathcal{SPN}^*$.

Proof.

1. For every reachable marking $\bar{\boldsymbol{m}}$ there is a firing sequence σ such that $\bar{\boldsymbol{m}}_0[\sigma > \bar{\boldsymbol{m}}$, i.e., $\bar{\boldsymbol{m}} - \bar{\boldsymbol{m}}_0 = \bar{\boldsymbol{A}}\bar{\boldsymbol{\sigma}}$. By combining Definition 12.1 with Definition 10.22 it follows that $\boldsymbol{\alpha}_{p_r^*} = \boldsymbol{b}_r = \boldsymbol{a}_r \boldsymbol{A}$. Combining these results for $p \in \mathcal{P}^*$ gives:

$$\bar{m}(p_r^*) - \bar{m}_0(p_r^*) = \boldsymbol{\alpha}_{p_r^*}\bar{\boldsymbol{\sigma}} = \boldsymbol{a}_r \boldsymbol{A}\bar{\boldsymbol{\sigma}} = \boldsymbol{a}_r(\boldsymbol{m} - \boldsymbol{m}_0).$$

   It should be noted that neither $\boldsymbol{a}_r$ nor σ is uniquely defined. However, for all $\boldsymbol{a}_r^1, \boldsymbol{a}_r^2$ satisfying the conditions in Definition 10.22 and all σ_i such that $\boldsymbol{m}_0[\sigma_i > \boldsymbol{m}, i \in \{1,2\}$, we have

$$\boldsymbol{a}_r^1 \boldsymbol{A}\bar{\boldsymbol{\sigma}}_1 = \boldsymbol{b}_r\bar{\boldsymbol{\sigma}}_1 = \boldsymbol{a}_r^2 \boldsymbol{A}\bar{\boldsymbol{\sigma}}_1 = \boldsymbol{a}_r^2(\boldsymbol{m} - \boldsymbol{m}_0) = \boldsymbol{a}_r^2 \boldsymbol{A}\bar{\boldsymbol{\sigma}}_2,$$

   so that the marking of the places $p \in \mathcal{P}^*$ is uniquely determined from the marking of the places $p \in P$, independent of the choice of $\boldsymbol{a}_r$ and firing sequence σ.

2. By construction of the bag count places, for every firing sequence σ from $\boldsymbol{m}_0$ to $\boldsymbol{m}$, for every bag $\boldsymbol{r}$, $\bar{m}(p_r^*) - \bar{m}_0(p_r^*)$ indicates exactly how many times bag $\boldsymbol{r}$ is deposited minus the number of times bag $\boldsymbol{r}$ is consumed. Part 1 of the proof

indicates that there is a unique difference $\bar{m}(p_r^*) - \bar{m}_0(p_r^*)$ corresponding to $\boldsymbol{m} - \boldsymbol{m}_0$. As a consequence, $\sum_{\boldsymbol{r}\in\mathcal{R}(T)}(\bar{m}(p_r^*) - \bar{m}_0(p_r^*))\boldsymbol{r}$ is independent of σ and thus $\boldsymbol{m}$ can be found by adding $\bar{m}(p_r^*) - \bar{m}_0(p_r^*)$ times bag $\boldsymbol{r}$ for every bag $\boldsymbol{r} \in \mathcal{R}(T)$ to the initial marking $\boldsymbol{m}_0$. □

Lemma 12.5. Let $\mathcal{SPN}$ be a structurally bounded $\mathcal{S}\Pi^2$-net and let $\mathcal{SPN}^*$ be its corresponding $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net. For every initial marking $\boldsymbol{m}_0$ of $\mathcal{SPN}$, an initial marking $\bar{\boldsymbol{m}}_0$ of $\mathcal{SPN}^*$ can be chosen such that $\bar{m}(p_r^*) \geq 1$, $\boldsymbol{r} \in \mathcal{R}(T)$, for all $\bar{\boldsymbol{m}} \in \mathcal{M}(\mathcal{SPN}^*, \bar{\boldsymbol{m}}_0)$.

Proof. Theorem 12.4 provides $\bar{m}(p_r^*) - \bar{m}_0(p_r^*) = \boldsymbol{a}_r(\boldsymbol{m} - \boldsymbol{m}_0)$ and since $(\mathcal{SPN}, \boldsymbol{m}_0)$ is bounded there is a constant C_p such that $0 \leq m(p) < C_p$ for all $p \in P$. Therefore

$$C_1 = \sum_{p\in P} \min(0, a_r(p)C_p) \leq \boldsymbol{a}_r\boldsymbol{m} \leq \sum_{p\in P} \max(0, a_r(p)C_p) = C_2,$$

so by taking initial marking $\bar{m}_0(p_r^*) = 1 - C_1 + \boldsymbol{a}_r\boldsymbol{m}_0$, we get

$$\bar{m}(p_r^*) = \bar{m}_0(p_r^*) + \boldsymbol{a}_r(\boldsymbol{m} - \boldsymbol{m}_0) = 1 - C_1 + \boldsymbol{a}_r\boldsymbol{m} \geq 1.$$ □

Lemma 12.6. The $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net $\mathcal{SPN}^*$ corresponding to a structurally bounded $\mathcal{S}\Pi^2$-net $\mathcal{SPN}$ is structurally bounded.

Proof. By Theorem 12.4, in $\mathcal{SPN}^*$ there is a one-to-one correspondence between the marking of the places $p \in P$ and the marking of the places $p \in \mathcal{P}^*$. Since $\mathcal{SPN}$ is bounded for every initial marking $\boldsymbol{m}_0$ and the marking of places $p \in \mathcal{P}^*$ is given by the linear equations (12.1), $\mathcal{SPN}^*$ is also bounded for every initial marking $\bar{\boldsymbol{m}}_0$. □

Lemma 12.7. Consider the $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net $\mathcal{SPN}^* = (\bar{P}, T, \bar{I}, \bar{O}, Q)$ of an $\mathcal{S}\Pi^2$-net $\mathcal{SPN}$. Removing all original places $p \in P$ from $\mathcal{SPN}^*$ and all arcs incident to the places $p \in P$ yields ℓ state machines: $\mathcal{SM}^1, \ldots, \mathcal{SM}^\ell$. Each $\mathcal{SM}^i$ corresponds to a common input bag class: $\mathcal{SM}^i = (\mathcal{P}^i, \mathcal{T}^i, I^i, O^i, Q^i)$, with $\mathcal{P}^i = \mathcal{P}(CI^i) \bigcap \mathcal{P}^*$, $\mathcal{T}^i = \mathcal{T}(CI^i)$, and where I^i, O^i, Q^i are $\bar{I}, \bar{O}, Q$ restricted to $\mathcal{P}^i$ and $\mathcal{T}^i$.

Proof. The proof follows by construction of the $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net. Every transition has exactly one bag count place in its input bag and exactly one bag count place in its output bag. Therefore, removing all original places from the net will yield a state machine. This state machine consists of ℓ separate components, because two bag count places p_1^* and p_2^* are connected in this state machine if and only if there is a CI-class CI^i such that $p_1^*, p_2^* \in \mathcal{P}(CI^i)$. □

Observe that marking $\boldsymbol{m}$ of $\mathcal{SPN}$ is characterized by the marking of the places $p \in \mathcal{P}^*$ in $\mathcal{SPN}^*$. Lemma 12.7 expresses that $\mathcal{SPN}^*$ without the original places yields ℓ state machines, one for each CI-class. We have the following interpretation of $\mathcal{S}\Pi^2$-nets: the marking $\boldsymbol{m}$ of an $\mathcal{S}\Pi^2$-net is characterized by the combination of the 'states' of each of its CI-classes, where the state of each CI-class is tracked by the marking of its state machine in the corresponding $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net.

Theorem 12.4 provides the interpretation of the $\boldsymbol{a}_r$-vectors. Every firing sequence in $\mathcal{SPN}$ which brings $\boldsymbol{m}_0$ to $\boldsymbol{m}$ is associated with a unique value for the difference in the number of times each bag $\boldsymbol{r}$ is deposited and consumed in the firing sequence. The vector $\boldsymbol{a}_r$ gives the transformation to calculate this number: $\boldsymbol{a}_r(\boldsymbol{m}-\boldsymbol{m}_0)$, that turns out to be independent of the firing sequence. Thus, the $\boldsymbol{a}_r$-vectors are used to track the 'state' of each of the CI-classes.

12.3 Decomposition

Building on the insights of the previous section, in this section we will decompose the equilibrium distribution of an $\mathcal{S}\Pi^2$ into a product of the invariant measures of the state machines corresponding to these CI-classes. In Theorem 11.10, the decomposition of an $\mathcal{S}\Pi$-net can be such that a subnet is formed by multiple connected common input bag classes. Here, we take Theorem 11.10 as the starting point to derive a decomposition result for $\mathcal{S}\Pi^2$-nets, which decomposes an $\mathcal{SPN}$ in all its common input bag classes.

Recall that in decomposition Theorem 11.10 two types of place sets play a key-role: the conflict place set and the surplus place set. Decomposition is established if the places in the intersection of those two sets can be removed from the net so that live components remain. Since in a $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$ the bag count places form a sufficient place set, the direct consequence is that the set of all original places forms a surplus place set, which implies that all conflict places can be assigned to be surplus places. This leads to the application of Theorem 11.10 in Theorem 12.8.

Note that a state machine Petri net is equivalent to a Jackson network, see also [456]. So, the routing chain of a state machine is equivalent to the well-known traffic equations from queueing theory. And since the structure of a state machine induces that each T-invariant has a closed support, with $\bar{\boldsymbol{m}}^i$ its marking, the equilibrium distribution of a state machine $\mathcal{SM}^i$ as introduced in Lemma 12.7 is as follows:

$$\pi^{SM}(\bar{\boldsymbol{m}}^i) = C \prod_{\boldsymbol{r}\in\mathcal{R}(\mathcal{T}^i)} y^i(\boldsymbol{r})^{\bar{\boldsymbol{m}}^i(p_r^*)} \qquad ,\bar{\boldsymbol{m}}^i \in \{\bar{\boldsymbol{m}}^i : \sum_{\boldsymbol{r}\in\mathcal{R}(\mathcal{T}^i)} \bar{\boldsymbol{m}}^i(p_r^*)\},$$

where $y^i(\cdot)$ is the solution of the routing chain (10.8) of state machine $\mathcal{SM}^i$, and C is a normalizing constant.

Theorem 12.8. Consider an $\mathcal{S}\Pi^2$-net $\mathcal{SPN}=(P,T,I,O,Q)$ with its $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net $\mathcal{SPN}^*$, a set of vectors $\boldsymbol{a}_r, \boldsymbol{r}\in\mathcal{R}(T)$ satisfying the conditions of Definition 10.22, and an initial marking $\bar{\boldsymbol{m}}_0$ satisfying the conditions of Lemma 12.3. Then, the equilibrium distribution π of $\mathcal{SPN}$ is equal to the equilibrium distribution $\bar{\pi}$ of $\mathcal{SPN}^*$, of which the invariant measure is a product over the invariant measures of the state machines:

$$\pi(\boldsymbol{m}) = \bar{\pi}(\bar{\boldsymbol{m}}) = B\prod_{i=1}^{\ell} \pi^{\mathcal{SM}^i}(\bar{\boldsymbol{m}}^i) \quad ,\boldsymbol{m}\in\mathcal{M}(\mathcal{SPN},\boldsymbol{m}_0), \tag{12.2}$$

Proof. By Lemma 12.7, removing all original places $p \in P$ from $\mathcal{SPN}^*$ yields ℓ state machines: $\mathcal{SM}^1, \ldots, \mathcal{SM}^\ell$; each $\mathcal{SM}^i$ corresponding to exactly one common input bag class. Next, we obtain from Theorem 12.4 that $\mathcal{P}^*$ is a sufficient place set. Therefore, the set of original places P is a surplus place set. By construction, all conflict places of a $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net are original places, i.e., $\mathcal{P}^{con} \subseteq P$. Since every transition is connected to a bag count place, no complete input bag is contained in the conflict place set, i.,e $\nexists t \in \bar{P}$ for which $\{p \in \bar{P} \mid \bar{I}_p(t) > 0\} \subset (\mathcal{P}^{con} \bigcap P)$. Theorem 11.10 and Lemma 12.6 complete the proof. □

12.4 Examples

This section illustrates the similarities and differences between Theorem 11.10 and Theorem 12.8 via three examples. The first example is an $\mathcal{S}\Pi^2$-net consisting of two CI-classes linked by a single conflict place. This conflict place will form a surplus place set by itself which means that both Theorem 11.10 and Theorem 12.8 give us the means to decompose it into two separate CI-classes. This example shows that both methods result in the same decomposition, however they follow a different path to obtain this decomposition. The second example is an $\mathcal{S}\Pi^2$-net, with three CI-classes, that can be decomposed in two ways into two parts using Theorem 11.10. Theorem 12.8, enables us to decompose it into three parts, one for each CI-class. The third example is an $\mathcal{S}\Pi^2$-net that has three CI-classes, where all places are conflict places. Obviously, Theorem 11.10 will not lead to a decomposition, whereas Theorem 12.8 again allows complete decomposition into CI-classes. This example shows that even if the CI-classes are strongly intertwined and the product form over the places does not seem to be able to be decomposed, it is still possible to separate the different CI-classes and identify their behavior separately. Finally, Example 12.12 is obtained from [270], and provides an illustration of Theorem 12.8 when a probabilistic output bag is involved.

Example 12.9. Consider the stochastic Petri net $\mathcal{SPN}$ displayed in Figure 12.1. From the incidence matrix

$$\boldsymbol{A} = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & -1 & 1 & -1 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 \end{pmatrix},$$

we obtain two minimal support T-invariants $\boldsymbol{x}^1 = (1100)$ and $\boldsymbol{x}^2 = (0011)$, and five minimal support P-invariants $\boldsymbol{y}^1 = (110000)$, $\boldsymbol{y}^2 = (101100)$, $\boldsymbol{y}^3 = (101010)$, $\boldsymbol{y}^4 = (000101)$ and $\boldsymbol{y}^5 = (000011)$ of which the first four are linearly independent. The two T-invariants are both closed and cover all transitions, so $\mathcal{SPN}$ is an $\mathcal{S}\Pi^2$-net. The T-invariants are not in common input bag relation, therefore $\mathcal{SPN}$ has

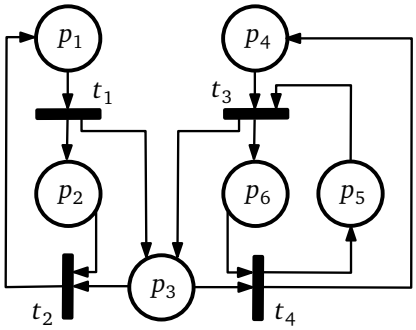


Figure 12.1: The $\mathcal{SPN}$ of Example 12.9.

two common input bag classes $CI^1 = \{\boldsymbol{x}^1\}$ and $CI^2 = \{\boldsymbol{x}^2\}$. This gives us one conflict place set $\{p_3\}$. Using the P-invariants we find that $\mathcal{P}^1 = \{p_2, p_3, p_5, p_6\}$ and $\mathcal{P}^2 = \{p_1, p_3, p_4, p_6\}$ are surplus place sets. Both these sets give $\mathcal{P}^{sur} \bigcap \mathcal{P}^{con} = \{p_3\}$, so in both cases Theorem 11.10 provides a decomposition into SPN^1 consisting of places $\{p_1, p_2\}$ and transitions $\{t_1, t_2\}$ and SPN^2 consisting of places $\{p_4, p_5, p_6\}$ and transitions $\{t_3, t_4\}$ (see Figure 12.2a). The equilibrium distribution of $\mathcal{SPN}$ is given by:

$$\begin{aligned}
\pi(\boldsymbol{m}) &= B\pi_y^{\mathcal{SPN}^1}(\boldsymbol{m}^1)\pi_y^{\mathcal{SPN}^2}(\boldsymbol{m}^2) \\
&= B\left(\frac{\mu_1}{\mu_2}\right)^{m(p_2)}\left(\frac{\mu_4}{\mu_3}\right)^{m(p_5)} && (12.3) \\
&= B\left(\frac{\mu_2}{\mu_1}\right)^{m(p_1)}\left(\frac{\mu_4}{\mu_3}\right)^{m(p_4)}, \quad \boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0), && (12.4)
\end{aligned}$$

where the form (12.3) is obtained when surplus place set $\mathcal{P}^1$ is used, and (12.4) when surplus place set $\mathcal{P}^2$ is used.

Now, let us apply Theorem 12.8. First we construct the $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net of $\mathcal{SPN}$ by adding four bag count places, $p_1^*, \ldots, p_4^*$. Now, removing the original places $p_1, \ldots, p_6$, gives the net shown in Figure 12.2b. This leads to the following equilibrium distribution, for $\boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)$:

$$\begin{aligned}
\pi(\boldsymbol{m}) &= B\pi^{\mathcal{SM}^1}(\bar{\boldsymbol{m}}^1)\pi^{\mathcal{SM}^2}(\bar{\boldsymbol{m}}^2) \\
&= B\left(\frac{1}{\mu_1}\right)^{\bar{m}(p_1^*)}\left(\frac{1}{\mu_2}\right)^{\bar{m}(p_2^*)}\left(\frac{1}{\mu_3}\right)^{\bar{m}(p_3^*)}\left(\frac{1}{\mu_4}\right)^{\bar{m}(p_4^*)} \\
&= B\left(\frac{1}{\mu_1}\right)^{\boldsymbol{a}_{I(t_1)}\boldsymbol{m}}\left(\frac{1}{\mu_2}\right)^{\boldsymbol{a}_{I(t_2)}\boldsymbol{m}}\left(\frac{1}{\mu_3}\right)^{\boldsymbol{a}_{I(t_3)}\boldsymbol{m}}\left(\frac{1}{\mu_4}\right)^{\boldsymbol{a}_{I(t_4)}\boldsymbol{m}}.
\end{aligned}$$

One of the possible choices for the vectors $\boldsymbol{a}_r$ is $\boldsymbol{a}_{I(t_1)} = (1,0,0,0,0,0)$ and $\boldsymbol{a}_{I(t_3)} = (0,0,0,1,0,0)$. This choice corresponds to (12.3), so to choosing $\mathcal{P}^{sur} = \mathcal{P}^1$ in

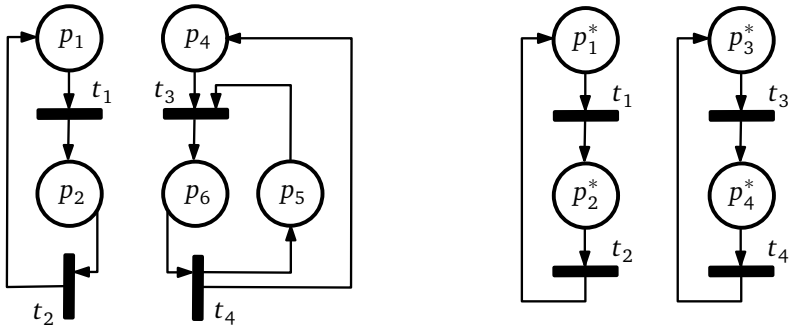


Figure 12.2: (a) Decomposition Ex. 12.9 via Thm. 11.10 (b) Decomposition via Thm. 12.8.

Theorem 11.10. A second possible choice is $\boldsymbol{a}_{I(t_1)} = (0,-1,0,0,0,0)$ and $\boldsymbol{a}_{I(t_3)} = (0,0,0,0,1,0)$, which corresponds to (12.4), and thus to choosing $\mathcal{P}^{sur} = \mathcal{P}^1$ in Theorem 11.10.

The first observation is that in this example Theorem 11.10 and Theorem 12.8 both lead to decomposition into two subnets and that the subnets correspond to the same parts of $\mathcal{SPN}$. However, the structure of the pieces is not necessarily the same. The subnet corresponding to CI^1 is the same in both cases, however the part corresponding to CI^2 has a different structure. The second observation is that a zero entries in all $\boldsymbol{a}_r$-vectors for a specific place $p \in P$, corresponds to assigning p as a surplus place. □

Example 12.10. Consider the $\mathcal{SPN}$ depicted in Figure 12.3a. From the incidence matrix:

$$\boldsymbol{A} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 1 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix},$$

we obtain the three minimal support T-invariants $\boldsymbol{x}^1 = (110000)$, $\boldsymbol{x}^2 = (01100)$ and $\boldsymbol{x}^3 = (000011)$ and four minimal support P-invariants $\boldsymbol{y}^1 = (1101000)$, $\boldsymbol{y}^2 = (1010010)$, $\boldsymbol{y}^3 = (0001100)$ and $\boldsymbol{y}^4 = (0000011)$, which are linearly independent. As the minimal support T-invariants are all closed and they cover all transitions, $\mathcal{SPN}$ is an $\mathcal{S}\Pi^2$-net. Furthermore, $\boldsymbol{x}^1$, $\boldsymbol{x}^2$ and $\boldsymbol{x}^3$ are not in common input bag relation so they result in three CI-classes, $CI^1 = \{\boldsymbol{x}^1\}$, $CI^2 = \{\boldsymbol{x}^2\}$ and $CI^3 = \{\boldsymbol{x}^3\}$. This results in the following conflict place set: $\mathcal{P}^{con} = \{p_2, p_3\}$.

Since the complete input bag of transition t_1 is contained in $\mathcal{P}^{con}$, Theorem 11.10 is not able to separate all CI-classes. However, both $\mathcal{P}^1 = \{p_2\}$ and $\mathcal{P}^2 = \{p_3\}$ are

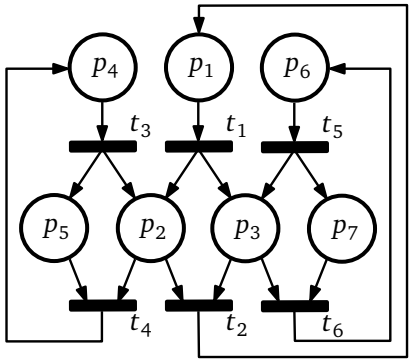
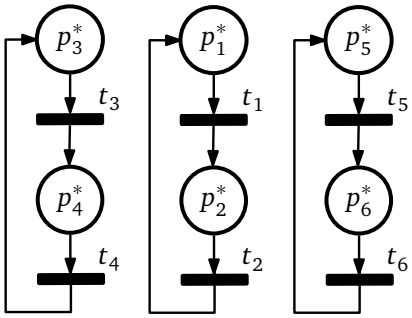


Figure 12.3: (a) $\mathcal{SPN}$ of Example 12.10



(b) Decomposition via Theorem 12.8.

surplus place sets. Both lead to a decomposition of the equilibrium distribution:

$$\pi(\boldsymbol{m}) = B\pi_y^{\mathcal{SPN}^1}(\boldsymbol{m}^1)\pi_y^{\mathcal{SPN}^2}(\boldsymbol{m}^2),$$

where in case of decomposition via $\mathcal{P}^1$ the two subnetworks are $\mathcal{SPN}^1 = \{CI(\boldsymbol{x}^1), CI(\boldsymbol{x}^3)\}$ and $\mathcal{SPN}^2 = \{CI(\boldsymbol{x}^2)\}$, while via $\mathcal{P}^2$ the two subnetworks are $\mathcal{SPN}^1 = \{CI(\boldsymbol{x}^1), CI(\boldsymbol{x}^2)\}$ and $\mathcal{SPN}^2 = \{CI(\boldsymbol{x}^3)\}$.

To illustrate the power of Theorem 12.8 over Theorem 11.10, we construct the $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net of $\mathcal{SPN}$. By adding the six bag count places, $p_1^*, \ldots, p_6^*$, to the net and then removing all original places, $p_1, \ldots, p_7$, we obtain the net shown in Figure 12.3b. A simple choice of the $\boldsymbol{a}_r$-vectors is allowed, similar to the previous example: $\boldsymbol{a}_{I(t_1)} = (1,0,0,0,0,0)$, $\boldsymbol{a}_{I(t_2)} = (-1,0,0,0,0,0)$, $\boldsymbol{a}_{I(t_3)} = (0,0,0,1,0,0,0)$, $\boldsymbol{a}_{I(t_4)} = (0,0,0,-1,0,0,0)$, $\boldsymbol{a}_{I(t_5)} = (0,0,0,0,0,1,0)$ and $\boldsymbol{a}_{I(t_6)} = (0,0,0,0,0,-1,0)$. This yields the following equilibrium distribution:

$$\begin{aligned}
\pi(\boldsymbol{m}) &= B\pi^{\mathcal{SM}^1}(\bar{\boldsymbol{m}}^1)\pi^{\mathcal{SM}}(\bar{\boldsymbol{m}}^2)\pi^{\mathcal{SM}}(\bar{\boldsymbol{m}}^3) \\
&= B\left(\frac{\mu_2}{\mu_1}\right)^{\bar{m}(p_1^*)}\left(\frac{\mu_4}{\mu_3}\right)^{\bar{m}(p_3^*)}\left(\frac{\mu_6}{\mu_5}\right)^{\bar{m}(p_5^*)} \\
&= B\left(\frac{\mu_2}{\mu_1}\right)^{m(p_1)}\left(\frac{\mu_4}{\mu_3}\right)^{m(p_4)}\left(\frac{\mu_6}{\mu_5}\right)^{m(p_6)}, \quad \boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0).
\end{aligned}$$

Thus, Theorem 12.8 enables a decomposition in the three cyclic state machines corresponding to the three CI-classes. □

Example 12.11. Consider the $\mathcal{SPN}$ of Figure 12.4a, with the following incidence matrix

$$\boldsymbol{A} = \begin{pmatrix} -1 & 1 & -1 & 1 & -1 & 1 \\ -1 & 1 & 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 \end{pmatrix},$$

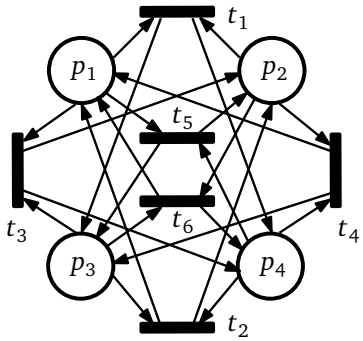
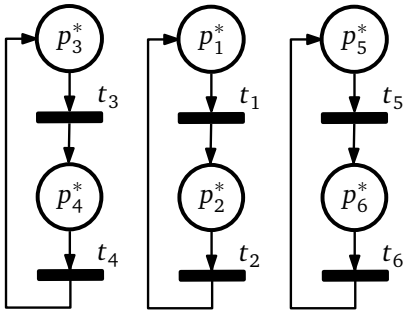


Figure 12.4: (a) $\mathcal{SPN}$ of Example 12.11



(b) Decomposition via Theorem 12.8.

There are three minimal support T-invariants $\boldsymbol{x}^1 = (110000)$, $\boldsymbol{x}^2 = (001100)$, $\boldsymbol{x}^3 = (000011)$, and one minimal support P-invariant $\boldsymbol{y}^1 = (1111)$. All the T-invariants are closed so it is an $\mathcal{S}\Pi^2$-net and none of the T-invariants are in common input bag relation, so there are three CI-classes, $CI^1 = \{\boldsymbol{x}^1\}$, $CI^2 = \{\boldsymbol{x}^2\}$ and $CI^3 = \{\boldsymbol{x}^3\}$. All places belong to each of the three CI-classes so the set of conflict places is $\{p_1, p_2, p_3, p_4\}$. Clearly, Theorem 11.10 does not lead to a decomposition. For Theorem 12.8, add the six bag count places to obtain the $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net with incidence matrix:

$$\bar{\boldsymbol{A}} = \begin{pmatrix} -1 & 1 & -1 & 1 & -1 & 1 \\ -1 & 1 & 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 \\ -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix},$$

A possible choice is for a set of vectors $\boldsymbol{a}_r, \boldsymbol{r} \in \mathcal{M}(T)$ is: $\boldsymbol{a}_{I(t_1)} = (1/2, 1/2, 0, 0)$, $\boldsymbol{a}_{I(t_2)} = (-1/2, -1/2, 0, 0)$, $\boldsymbol{a}_{I(t_3)} = (1/2, 0, 1/2, 0)$, $\boldsymbol{a}_{I(t_4)} = (-1/2, 0, -1/2, 0)$, $\boldsymbol{a}_{I(t_5)} = (0, -1/2, -1/2, 0)$, and $\boldsymbol{a}_{I(t_6)} = (0, 1/2, 1/2, 0)$.

By removing the original places from the Petri net we obtain the net shown in Figure 12.4b. Note that this net is the same as the reduced net we obtained in Example 12.10. Thus, we obtain the following equilibrium distribution, for $\boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)$:

$$\begin{aligned} \pi(\boldsymbol{m}) &= B\pi^{\mathcal{SM}^1}(\bar{\boldsymbol{m}}^1)\pi^{\mathcal{SM}^2}(\bar{\boldsymbol{m}}^2)\pi^{\mathcal{SM}^3}(\bar{\boldsymbol{m}}^3) \\ &= B\left(\frac{\mu_2}{\mu_1}\right)^{\bar{m}(p_1^*)}\left(\frac{\mu_4}{\mu_3}\right)^{\bar{m}(p_3^*)}\left(\frac{\mu_6}{\mu_5}\right)^{\bar{m}(p_5^*)} \end{aligned}$$

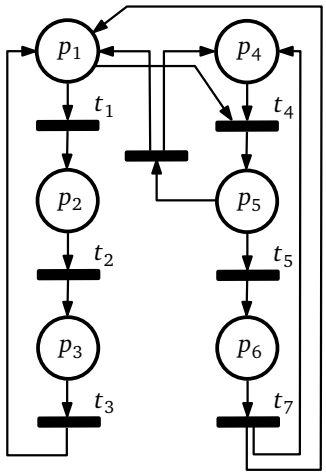
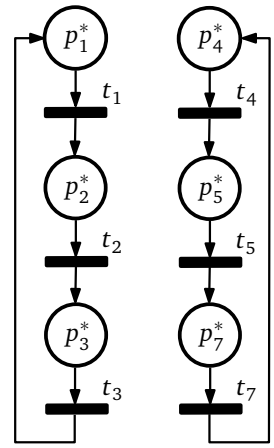


Figure 12.5: (a) $\mathcal{SPN}$ of Example 12.12



(b) Decomposition via Theorem 12.8.

$$= B\left(\frac{\mu_2}{\mu_1}\right)^{\frac{1}{2}(m(p_1)+m(p_2))}\left(\frac{\mu_4}{\mu_3}\right)^{\frac{1}{2}(m(p_1)+m(p_3))}\left(\frac{\mu_5}{\mu_6}\right)^{\frac{1}{2}(m(p_2)+m(p_3))}.$$

□

Example 12.12. Consider the $\mathcal{S}\Pi^2$-net of Figure 12.5a, taken from [270], which has minimal minimal closed support T-invariants $\boldsymbol{x}^1 = (111000), \boldsymbol{x}^2 = (0001100)$ and $\boldsymbol{x}^3 = (0001011)$. The CI-classes are: $CI^1 = \{\boldsymbol{x}^1\}$ and $CI^2 = \{\boldsymbol{x}^2, \boldsymbol{x}^3\}$. Theorem 11.10 does not provide a decomposition, since it would require the removal of the complete input bag of transition t_1. Since t_5 are t_6 have the same input bag, the probabilistic output bag transformation is applied, and Theorem 12.8 requires the creation of only six bag count places. The decomposed net is shown in Figure 12.5b. A possible choice for the $\boldsymbol{a}_r$-vectors is (also see [270]): $\boldsymbol{a}_{I(t_1)} = (0,-1,-1,0,0,0)$, $\boldsymbol{a}_{I(t_2)} = (0,1,0,0,0,0)$, $\boldsymbol{a}_{I(t_3)} = (0,0,1,0,0,0)$, $\boldsymbol{a}_{I(t_4)} = (0,0,0,1,0,0)$, $\boldsymbol{a}_{I(t_5)} = (0,0,0,0,1,0)$, and $\boldsymbol{a}_{I(t_7)} = (0,0,0,0,0,1)$, which leads to the following equilibrium distribution, for $\boldsymbol{m} \in \mathcal{M}(\mathcal{SPN}, \boldsymbol{m}_0)$:

$$\begin{aligned}\pi(\boldsymbol{m}) &= B\pi^{\mathcal{SM}^1}(\bar{\boldsymbol{m}}^1)\pi^{\mathcal{SM}^2}(\bar{\boldsymbol{m}}^2)\\ &= B\left(\frac{\mu_1}{\mu_2}\right)^{m(p_2)}\left(\frac{\mu_1}{\mu_3}\right)^{m(p_3)}\left(\frac{1}{\mu_4}\right)^{m(p_4)}\left(\frac{1}{\mu_5}\right)^{m(p_5)}\left(\frac{\mu_6}{(\mu_5+\mu_6)\mu_7}\right)^{m(p_6)}\end{aligned}$$

□

CHAPTER 13

Petri Nets in Practice

13.1 Introduction

Understanding the behavior of dynamic systems is often difficult due to complex causal relationships of the system elements involved. Petri nets provide a uniform language for modeling and analysis, by which the design and operations of discrete event systems can be supported [25, 417, 657]. The particular features of Petri nets fit well to system characteristics that are prevalent in healthcare environments. Examples of such characteristics are competition over shared resources, synchronization of events, and parallelism of processes. Petri nets are therefore a promising modeling tool to accurately capture the complex patient flow dynamics of healthcare organizations. Another advantage of the Petri net language in applying it to healthcare processes is its twofold nature. Its graphical nature can be used to visually demonstrate the behavior of a system, so that it can serve as a communication medium between researchers and practitioners. Its mathematical nature makes it possible to formally define the behavior of the system, so that it can serve as a tool for performance analysis.

Petri nets have been widely applied in industrial areas, such as communication and computer systems (e.g., [50, 54, 178]), manufacturing systems (e.g., [159, 204, 387]), and supply chains (e.g., [183, 331, 610]). Application of Petri nets to healthcare logistics can also be found. Many of these contributions concern the use of Petri nets in designing workflow management systems (e.g., [409, 512, 580]), i.e., computer systems that support streamlined execution of operational processes by defining and managing the series of tasks involved. Stochastic Petri net studies aimed at performance analysis are also available. These contributions generally apply computer simulation to obtain numerical results; some take a care chain perspective [316, 358], the majority considers single departments, such as emergency departments [14, 121, 141, 649], inpatient care services [391, 493], or diagnostics facilities [554]. The advantage of simulation approaches is their flexibility and therefore modeling power. However, disadvantages are that model construction and evaluation can be very time-consuming and that the nature of simulation studies is typically context specific, which limits the generalizability of their application and findings.

A generic analytical stochastic Petri net framework would be a valuable addition to the existing literature. By developing such a framework, one can accomplish high-level insight in the behavior of complex care chains and efficient computation of relevant performance measures, thereby supporting decision makers in selecting optimal design alternatives and operational policies. It will be suitable to address strategic and tactical resource capacity planning and control decisions, such as service mix, case mix, capacity dimensioning, capacity allocation, and admission control (see Chapter 2 for an elaborate description of these decisions). The results from Chapters 10–12 contribute to the body of knowledge on analytic performance analysis that is computationally efficient and that provides insight in the behavior of parts of the system without having to consider the complete system. Since the status of this ongoing research is anything but a ready-to-use decision support tool, in the current chapter we provide the bridge between Petri net theory and its application in healthcare.

The chapter is organized as follows. Section 13.2 provides a summary of the theoretical results of Chapters 10–12. A decision support tool based on stochastic Petri nets requires both a model construction component and a performance analysis component. Section 13.3 sketches actions involved in constructing abstract models representing given practical situations. Finally, in Section 13.4 we describe research directions to obtain increased modeling power, and to enable approximative performance analysis for 'practical Petri nets' which in general do not allow for exact analytic computation of performance measures.

13.2 Results overview

In Chapters 10–12, we have surveyed, unified and extended structural product form and decomposition results for stochastic Petri nets. Group-local-balance has been shown to be the unifying concept between known product form results for stochastic Petri nets and has provided the ground to formulate necessary and sufficient structural conditions for product form and decomposition and to obtain a structural and intuitive explanation of these conditions, completely in terms of P- and T-invariants. Product form has been discussed in Chapter 10 and decomposition was addressed in Chapters 11 and 12. Below we provide an overview of the main results.

Theorem 10.5 opens the batch-routing queueing network literature for stochastic Petri nets as it provides the translation of product form results for batch routing queueing networks based on group-local-balance to stochastic Petri nets. Group-local-local balance implies that for product form a positive solution is required to the routing chain (10.8). Theorem 10.11 states that for a stochastic Petri net a positive solution for the routing chain exists if and only if it is an $\mathcal{S}\Pi$-net. Theorem 10.19 states that an $\mathcal{S}\Pi$-net has an equilibrium distribution that is a product form over the places of the network if and only if it satisfies group-local-balance. As such, Theorem 10.19 closes the cycle to batch-routing queueing networks. This brings us in the position to investigate the Petri net structure behind group-local-balance.

From Theorem 10.19 it appears that, in general, for group-local-balance to hold in an $\mathcal{S}\Pi$-net, an additional condition on the numerical values of the transition rates is required to be satisfied (see Lemma 10.18). Theorem 10.21 shows that for each minimal *closed* support T-invariant this numerical condition is satisfied irrespective of the numerical values of the transition rates. Therefore, for an $S\Pi$-net in which each minimal support T-invariant is a minimal *closed* support T-invariant, group-local-balance is satisfied, and thus product form holds.

In this way, we have unified the key steps presented in literature with respect to structural results for product form stochastic Petri nets. Henderson et al. [296] introduced the routing chain. Assuming that a positive solution exists to the global balance equations of the routing chain, they showed that if a closed form solution to ratio condition (10.9) on the solution of the routing chain can be found, this is the equilibrium distribution. Coleman et al. [130] identified the numerical condition, which we have stated in Lemma 10.18, under which such a closed form solution exists and is of product form. We have shown that both the results of Henderson et al. and Coleman et al. can be explained as originating from group-local-balance. The last step was to unify Theorem 10.21 with the characterization by Haddad et al. [270] and Mairesse et al. [406] of rate-insensitive product form stochastic Petri nets. Their algebraic definitions of respectively $S\Pi^2$-nets and deficiency zero $\mathcal{S}\Pi$-nets, subclasses of $S\Pi$-nets, were in Theorem 10.23 shown to be equivalent with our characterization of rate-insensitive product form stochastic Petri nets; Theorem 10.23 states that an $S\Pi$-net is an $S\Pi^2$-net if and only if *all* minimal support T-invariants are minimal *closed* support T-invariants.

Product form results for network structures often allow for hierarchical composition and decomposition of subnetworks. When interested in global characteristics of a network it is convenient to decompose the network so that local characteristics can be investigated without considering the complete network in detail. Chapter 11 introduced decomposition results by which subnetworks can be identified in which a given product form stochastic Petri net can be decomposed. These subnetworks correspond to one or more common input bag classes, equivalence classes of minimal closed support T-invariants connected by having an input bag in common. Essential in achieving the decomposition is the notion of the sufficient place set of a Petri net, the set of places sufficient for uniquely characterizing the marking of a Petri net at all its places. The complement of the sufficient place set is the surplus place set, places that can be omitted in characterizing the marking of the Petri net. A procedure to characterize surplus place sets of a Petri net from its P-invariants is provided in Lemma 11.3. Removing conflict places that can be assigned as a surplus place yields decomposition. The restriction is that no complete input bag may be removed. To be specific, Theorem 11.10 states that if a sufficient place set can be found so that there is no input bag of which all places are both surplus and conflict places, a product form stochastic Petri net decomposes into subnets each corresponding to one or more common input bag classes. The steps that have to be performed to verify and construct product form and to obtain all possible decompositions are summarized in Algorithms 10.27 and 11.13.

Building on decomposition Theorem 11.10 and the characterization of $\mathcal{S}\Pi^2$-nets by Haddad et al. [270] (in thesis formulated in Definition 10.22), in Theorem 12.8 we presented an additional decomposition result. Starting from an arbitrary $\mathcal{S}\Pi^2$-net, and adding 'bag count places', Definition 12.1 introduced the Bag-Count-Place-Extended Petri net of an $\mathcal{S}\Pi^2$-net ($\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net). Theorem 12.4 shows that the algebraic characterization of $\mathcal{S}\Pi^2$-nets of [270] induces a one-to-one correspondence between the marking of the original places and the additionally constructed bag count places. This one-to-one correspondence implies that the bag count places of a $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$ form a sufficient place set, and thus that the equilibrium distribution of the bag count places provides an equilibrium distribution of the original places. In addition, by construction the bag count places of a $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net are non-conflict places. These observations enabled us to apply decomposition Theorem 11.10 to the $\mathcal{BCPE}$-$\mathcal{S}\Pi^2$-net: Theorem 12.8 states that each rate-insensitive product form stochastic Petri net decomposes into subnets which correspond to exactly one common input bag class. Lemma 12.7 shows that the bag count place description is such that a state machine is associated with each of the common input bag classes of an $\mathcal{S}\Pi^2$-net. As a consequence, we have revealed the intuition that the behavior of an $\mathcal{S}\Pi^2$-net is a result of a complex interaction between underlying state machines per common input bag class.

Finally, observe that characterizing product form for a stochastic Petri net can be done completely in terms of its *T*-invariants, while decomposition of the network into subnetworks not only requires the *T*-invariants, but also its *P*-invariants.

13.3 Care chain modeling

A prerequisite for improving healthcare via performance analysis with stochastic Petri nets, is the ability to design a Petri net representation of a healthcare delivery system. Only after an adequate model is specified, it can be analyzed to learn about and improve the behavior of the original system [360]. Identifying the adequate level of abstraction for such a model is not a trivial task. Depending on the research purpose and scope, the most relevant system characteristics need to be identified and incorporated in a model [417]. Also, since developing Petri net models for large scale systems is complex, time-consuming, and requires a great deal of expertise, in the intended healthcare decision support framework, model construction is preferably (partly) automated. Finally, in general, the complexity of real-life systems prohibits analytical calculation of relevant performance measures. Therefore, it is worthwhile to take this issue into account during the model construction phase, so that the constructed model structures resemble model structures for which analytical results can be obtained.

Petri nets provide a language to model discrete event systems. A discrete event system involves a chronological sequence of events, where each event occurs at an instant in time and marks a change of state in the system [383]. Healthcare delivery systems can be seen as discrete event systems, where the system state is described by the status of each patient and each resource [141, 360]. During the

operation of the healthcare delivery system, the system moves from one state to the next when actions are performed, e.g., patient admissions and discharges. The set of possible actions in a certain system state define the set of 'enabled events'. In general, resources are required to perform actions, and when events happen, new events are enabled. Building a Petri model representing a practical healthcare environment thus requires the translation of real-world system information, such as possible patient pathways, resource requirements, and resource availability, into an abstract model.

Manual collection of practical data and creation of the Petri net model is very time-consuming [141]. Moreover, the quality of the model is in this case dependent on the expertise level of the modeler [657]. In today's healthcare information systems, for instance via electronic medical records, large amounts of historical data are available on the execution of the healthcare delivery process. Patients generate a sequence of digital messages during their clinical course, so-called 'event-logs'. The research field of 'process mining' focuses on exploiting these event logs to construct a model representation of a real-world system [581, 584]. Rather than for performance evaluation, process mining is originally used to obtain insight in how a process is performed when no formal documentation of a process exists, or to find out whether a process is performed according to the documented protocols. Since model representations obtained by process mining are often in the form of Petri nets (e.g., [358, 408, 409]), combining process mining and analytic performance evaluation via stochastic Petri nets appears promising to us.

Let us now provide some direction on the type the stochastic Petri net healthcare models that could support strategic and tactical decision making. A sample Petri net model is displayed in Figure 13.1. The example only explicitly contains one patient type and a selection of involved resources within the scope of a hospital; it contains the key-elements to be able to explain our main ideas. In a real-world example, more resources will be involved, and many more patient pathways will exist of which the transitions are in a similar way connected to the places modeling resources. Two type of tokens and places are used. *Patient places* are only marked by *patient tokens*, and *resource places* only by *resource tokens*. The marking of a token on a specific place represents the status of the corresponding patient or resource. The patient places represent stages in patient care pathways, and the resource places indicate whether resources are in use, ready for use, or unavailable. The interaction between patients and resources is modeled via the transitions, which represent actions that are performed in the patient treatment process. The example shows the idea of separating local and global behavior: the global behavior is a result of the complex interaction between individual parts formed by sets of patient pathways. Note that the example is not an $\mathcal{S}\Pi$-net, this issue will be addressed in Section 13.4.

For the patient type that is displayed in Figure 13.1 there exist three possible pathways: (1) Home – Outpatient clinic – Home, (2) Home – Outpatient clinic – CT + Lab – Outpatient clinic – Home, and (3) Home – Outpatient clinic – CT + Lab – Outpatient clinic – Operating theater – Recovery room – Inpatient clinic – Home. These three pathways are reflected by T-invariants $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$, with the supports

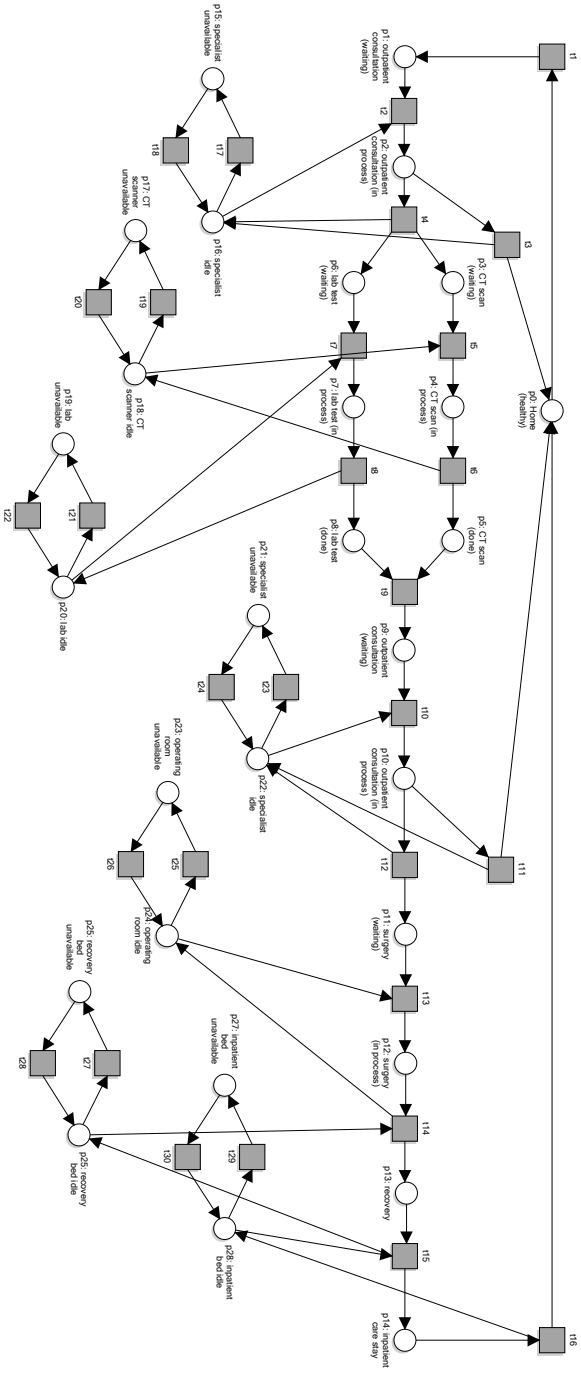


Figure 13.1: Sample Petri net.

$\|\boldsymbol{x}_1\| = \{t_1, t_2, t_3\}$, $\|\boldsymbol{x}_2\| = \{t_1, t_2, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}, t_{11}\}$, and $\|\boldsymbol{x}_3\| = \{t_1, t_2, t_4, t_5, t_6, t_7, t_8, t_9, t_{10}, t_{12}, t_{13}, t_{14}, t_{15}, t_{16}\}$. The initial marking only marks patient place p_1, reflecting the potential patient population, and the resource places p_{16}, p_{18}, p_{20}, p_{22}, p_{24}, p_{26}, and p_{28}, reflecting that all resources are initially available. Resource tokens are consumed when patient actions involving that resource are started, and these tokens are deposited when these action are finished.

The expressiveness of the Petri net language is illustrated at various points in this simple example. The competition mechanism over resources between patient pathways is for example reflected by places p_{11}, p_{24} and transition t_{13}: only one patient at a time can undergo surgery in an operating room. Transitions t_5, t_6, t_7, t_8 form an example where a healthcare system contains processes that can be performed in parallel: for this patient type there is no mandatory sequence in which the lab test and the CT scan have to be performed. Finally, places p_5, p_8 and transition t_9 show an example of synchronization: only when both the results of the CT scan and the lab test are available, the second outpatient consultation can take place.

13.4 Future research directions

Petri nets describing real-world organizations are not expected to be $\mathcal{S}\Pi$-nets. Since the results of Chapters 10–12 are restricted to $\mathcal{S}\Pi$-nets, developing a Petri net healthcare decision support framework clearly requires further research. The goal of this section is to describe several promising directions for future research.

The most prominent issue to address is that, like the example presented in the previous section, practical Petri nets do often not satisfy the underlying conditions for product form equilibrium distributions to hold. Exact performance analysis of healthcare organization via closed-form expressions is therefore in general not directly possible. Nevertheless, the obtained product form and decomposition results do provide directions for future research focused on approximative analysis. The structural characterization of product form that have been presented in terms of T-invariants, can help to identify which particular system characteristics destroy the product form property. Therefore, also for non-product form nets, product form analysis is of interest as it can provide directions for modifications by which the possibility of performance analysis via closed-form expression can be regained. Modifications to think of are to insert additional transitions so that a general Petri net becomes an $\mathcal{S}\Pi^2$-net, such as suggested in [269], or to add additional arcs to existing transitions to transform non-closed support T-invariants into closed support T-invariants. Inspiration for possible modifications might be obtained from [582], in which by modifying the original system approximations of queueing networks without product form solutions were proposed.

Obviously, modified Petri nets will not show the same behavior as the original nets. It is then interesting to gain insight into the impact the modifications have on system performance, to identify how well the modified systems approximate the original system. To this end, efforts can be put into the derivation of theoretical error bounds on the performance deviation between the original and modified systems. If

the size of the original net is not too large, another option is to compare the performance of the modified net to that of the original net by building a simulation model of the stochastic Petri net (using an available software tool such as CPN Tools [334] Yasper [586]), or numerically solving the associated continuous time Markov chain (for example with GreatSPN [419]). In conclusion, establishing general modification schemes for approximative analysis is an interesting research direction.

Motivated by the results of Henderson and Lucic [294, 295], we state that another promising direction to be able to analyze non-product form nets is to explore the concept of 'insensitivity'. Insensitivity in stochastic Petri nets induces that when (a selection of) the negative exponentially distributed firing times are replaced by generally distributed firing times, the equilibrium distribution of the net remains unchanged [293]. Henderson and Lucic presented several examples of stochastic Petri nets that allow generally distributed firing times for a selection of transitions without the product form property being destroyed. They showed that such insensitive transitions can in particular cases be considered as the aggregation of an underlying subnet. The other way around, by transition merging and marking amalgamation, non-product form nets were transformed into product form nets. An open challenge is to structurally characterize such insensitive stochastic Petri nets, and to formulate an algorithmic aggregation/disaggregation procedure by which product form results can be used to analyze non-product form nets. The interpretation of $\mathcal{S}\Pi^2$-nets in terms of underlying state machines per common input bag class, as formulated in Lemma 12.7, provides a good starting point to address this challenge.

With respect to decomposition results for non-product form nets, observe that although it is not an $\mathcal{S}\Pi$-net, the example in Figure 13.1 still reflects the key ideas behind the product form and decomposition results from Chapter 10–12. For instance, patient pathways with similar resource requirements are clustered by a set of T-invariants sharing an input bag, e.g., in our sample net T-invariants $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$ share input bag $\boldsymbol{I}(t_1)$. As such, a similar concept to, although not yet formally defined, common input bag classes still seems to exist. The places where these equivalence classes (to be formally defined) overlap, can again be seen as conflict places. Observe that the net is such that resource places can be defined which are both surplus and conflict places (e.g., places p_{16}, p_{18}, and p_{20}). It would therefore be worthwhile to explore whether it is possible to derive adjusted versions of our decomposition theorems, by which also these practical nets can be decomposed.

Another theoretical extension to perform is to include marking-dependent firing rates. Marking-dependent firing rates make it possible to let the firing rates depend on the number of patients present and the number of resources available. As such, their inclusion is indispensable when analyzing real-world healthcare systems, since for example more patients can be seen at the outpatient clinic when more doctors are available. In addition, it opens the opportunity to investigate tactical decision making. As an illustration, by letting firing rates depend on the length of the waiting lists per patient type, the performance of different admission control policies can be evaluated. The same holds for capacity allocation decisions, when for example firing rates are adjusted so that the resource share allocation to a certain patient

group increases when a relative large number of patients of that group are in the system. Since the analysis of stochastic Petri nets with marking-dependent firing rates relies heavily on the analysis of nets with marking-independent firing rates, in this thesis we restricted ourselves to marking-independent firing rates given by equation (10.6). Henderson et al. [296] and Haddad et al. [270] have described particular forms of marking-dependent transition rates under which the product form property is not destroyed. Further research is required to investigate for what forms of marking-dependent transition rates our decomposition results still hold.

A final issue to address is that of formulating and calculating performance measures based on the steady state probabilities of marking occurrences. Examples of such performance measures of interest are system throughput (i.e., the number of token arrivals at selected places during a time unit) or utilization of resources (i.e., the fraction of time during which selected resource tokens are in use). Some performance can be calculated directly via the closed-form expression of the equilibrium distribution, others require the computation of the normalizing constant. The major difficulty of straightforward calculation of the normalizing constant is the need to generate the reachability set, of which the size increases exponentially with both the number of tokens in the initial marking and the number of places. Various articles have introduced methods for efficient computation of performance measures (e.g., [25, 130, 524, 525]). Analogous to what has been done for product form queueing networks, these references formulated mean value and convolution algorithms that use recursive relations between local and global characteristics of stochastic Petri nets to calculate performance measures directly or via the normalization constant. The structural characterization of product form and decomposition in term of T- and P-invariants we provided, and the interpretation of $\mathcal{S}\Pi^2$-nets in terms of underlying state machines per common input bag class, open the opportunity to improve upon these algorithms.

Epilogue

This dissertation addressed the application of operations research techniques to the managerial field of healthcare resource capacity planning and control. Focusing on various design and organization issues for different types of care services, it emphasized the value of taking an integral perspective on logistical decision making. Creating alignment between strategic, tactical, and operational decisions, and facilitating coordination between the actors within a care chain, is demonstrated to improve performance on both quality and efficiency dimensions. In this concluding chapter, we discuss the lessons learnt with respect to the implementation of mathematical results, or better, with respect to the realization of practical impact.

Healthcare organizations are increasingly aware that the complexity of present-day healthcare delivery is such that outstanding medical knowledge by itself is not enough for care providers to be successful. Incorporating logistical knowledge is considered of growing importance in effectively managing interactions between patient pathways, competition over resources, and conflicting goals of stakeholders. The first practical impact of the research documented in this thesis is that is helps in creating the awareness among healthcare administrators for the positive effects of taking a systems view on healthcare delivery.

Describing the potential of a system-oriented approach started with the presentation of a taxonomy in Part II, which outlined what explicit planning decisions are involved in setting goals and deciding in advance what to do, how to do it, when to do it and who should do it. By discussing the trade-offs involved and making the interrelation between decisions explicit, it provided insights on how to optimize healthcare processes and on how suboptimization can be avoided. These insights were used in Part III and IV, which quantitatively illustrated the value of integrating capacity allocation, admission control, and appointment scheduling decisions in ambulatory care services. Part V did accordingly for case mix, surgical block scheduling, care unit partitioning, care unit size, and staff-shift scheduling decisions in surgical and inpatient care services. Part VI laid a theoretical foundation for a decision support tool by which the interrelation of service mix, case mix, capacity dimensioning, capacity allocation, and admission control in entire care chains can be studied.

Creating awareness for promising new ways of working is thus identified as the first strength of mathematical modeling. Creating awareness is an essential precondition for implementation of operations research models. We consider a definition of implementation stating that it is only realized when models are deployed in the actual planning and scheduling of operations, as being too tight. We would rather

stretch its definition to realizing impact. After creating awareness, impact is realized in three steps that are inseparably linked and that do not happen strictly chronologically: the modeling exercise, the calculations, and the decision-making process.

Realizing impact starts from the very first moment the modeling exercise takes off. Model building helps in posing questions concerning the system that have never been asked before, and forces answers to be formulated. As such, one obtains an improved understanding of the system under study and of who is when responsible for which decision. In addition, model building provides an incentive for collecting, cleansing, and organizing data. Giving practitioners insight in data on their care delivery realization, often results in direct improvement actions. These advantages of the modeling exercise have also been experienced and expressed by the clinicians involved in the various case studies presented.

The second step in realizing impact involves the actual calculations, based on which explicit recommendations to decision makers are formulated. Quantifying the effect via mathematical models offers the opportunity to investigate the effect of different alternatives in situations where actual experiments are impossible, because that is too costly, time-consuming, risky or unethical (e.g., the redesign of inpatient care services in Chapters 7 and 8), or because the problem concerns a future situation (e.g., the introduction of treatment plans in Chapters 5 and 6). In addition, employing mathematical modeling can provide solutions to planning issues that would otherwise remain uncovered due to their complexity (e.g., the development of day schedules in Chapter 4).

The third step concerns the ability of mathematical models to act as a communication tool in the eventual decision making process. The modeling phase provided insight into who the stakeholders are and which interests they have. The calculations showed the trade-offs that have to be taken into account in the act of balancing the interests of these stakeholders. As such, a shared understanding about the problem and possible solutions is created. Thereby, it facilitates the decision-making process by establishing increased mutual understanding between the actors. This can for example be crucial in successfully coming to agreements in the negotiation process of setting target patient performance indicators, in relation with associated resource utilization (e.g., acceptable access times for outpatients in Chapter 3, or rejection probability of inpatients in Chapter 7), or in relation with necessary sacrifices to staff preferences (e.g., not having the team meeting at a fixed time, or the acceptance of gaps in the clinician schedules in Chapter 4). By explicitly quantifying the consequences of different choices, the nature of the debate changes, as it becomes less emotional, and more rational, This makes a difference, especially in politically charged settings where large interests are at stake (such as in restructuring inpatient care services or changing a surgical block schedule). Thus, we claim that it is the complete course of building and deploying mathematical models that helps decision makers improving healthcare delivery.

We identified the potential of interconnecting the fields of medicine and applied mathematics. To further exploit this potential, involving more disciplines will be beneficial. When it concerns mathematical models to support daily operations,

knowledge from computer sciences helps transforming the models into software tools that are integrated with an organization's existing electronic database system (e.g., to support the appointment scheduling in the settings of Chapter 4 and 5). Next, in a human-oriented environment healthcare is, models should involve people and not exclude them. Translating modeling results to decision makers in a way that medical professionals keep their autonomy in the care delivery process is essential. Bringing in knowledge from social sciences can help communicating the insights and conclusions obtained from operations research models to problem owners in the most constructive way.

In conclusion, this thesis demonstrated that Operations Research can play an essential role in addressing the tough logistical challenges healthcare organizations face. Mathematical modeling can make a positive contribution to the achievement of higher quality and increased productivity of labor and capital. We are convinced that healthcare organizations can benefit from giving mathematical modeling a permanent position in their decision-making processes. Because implementation of solutions often requires people to do things differently, it often meets with resistance. A prerequisite for successful implementation of results is that of operations researchers and practitioners working closely together. This thesis intends to build a bridge between science and practice.

Bibliography

[1] E.H.L. Aarts and J.K. Lenstra. *Local search in combinatorial optimization*. Princeton University Press, Princeton, NJ, USA, 2003.

[2] W.J. Abernathy and J.C. Hershey. A spatial-allocation model for regional health-services planning. *Operations Research*, 20(3):629–642, 1972.

[3] Academic Medical Center. Brochure AMC. *Retrieved October 13, 2012, from http://issuu.com/amcamsterdam/docs/amcbrochure*, 2010.

[4] Academic Medical Center. Strategy statement 2011-2015. *Internal report*, 2011.

[5] Academic Medical Center. Annual report 2011. *Retrieved October 13, 2012, from http://www.amc.nl*, 2012.

[6] I.J.B.F. Adan, J.A. Bekkers, N.P. Dellaert, J. Jeunet, and J.M.H. Vissers. Improving operational effectiveness of tactical master plans for emergency and elective patients under stochastic demand and capacitated resources. *European Journal of Operational Research*, 213(1):290–308, 2011.

[7] I.J.B.F. Adan, J.A. Bekkers, N.P. Dellaert, J.M.H. Vissers, and X. Yu. Patient mix optimisation and stochastic resource requirements: A case study in cardiothoracic surgery planning. *Health Care Management Science*, 12(2):129–141, 2009.

[8] I.J.B.F. Adan, J.S.H. van Leeuwaarden, and E.M.M. Winands. On the application of Rouché's theorem in queueing theory. *Operations Research Letters*, 34(3):355–360, 2006.

[9] L.H. Aiken, S.P. Clarke, D.M. Sloane, J. Sochalski, and J.H. Silber. Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *Journal of the American Medical Association*, 288(16):1987–1993, 2002.

[10] L.H. Aiken, W. Sermeus, K. van den Heede, D.M. Sloane, R. Busse, M. McKee, L. Bruyneel, A.M. Rafferty, P. Griffiths, M.T. Moreno-Casbas, C. Tishelman, A. Scott, T. Brzostek, J. Kinnunen, R. Schwendimann, M. Heinen, D. Zikos, I. Strømseng Sjetne, H.L. Smith, and A. Kutney-Lee. Patient safety, satisfaction, and quality of hospital care: cross sectional surveys of nurses and patients in 12 countries in Europe and the United States. *British Medical Journal*, 344(3):1–14, 2012.

[11] L.H. Aiken, D.M. Sloane, J.P. Cimiotti, S.P. Clarke, L. Flynn, J.A. Seago, J. Spetz, and H.L. Smith. Implications of the California nurse staffing mandate for other states. *Health Services Research*, 45(4):904–921, 2010.

[12] E. Akcali, M.J. Côté, and C. Lin. A network flow approach to optimizing hospital bed capacity decisions. *Health Care Management Science*, 9(4):391–404, 2006.

[13] R. Akkerman and M. Knip. Reallocation of beds to reduce waiting time for cardiac surgery. *Health Care Management Science*, 7(2):119–126, 2004.

[14] G. Amodio, M.P. Fanti, L. Martino, A.M. Mangini, and W. Ukovich. A Petri net model for performance evaluation and management of emergency cardiology departments. In *Proceedings of the 35th International Conference on Operational Research Applied to Health Services*, Leuven, 2009.

[15] D.R. Anderson, D.J. Sweeney, T.A. Williams, and M. Wisniewski. *An introduction to management science: quantitative approaches to decision making*. Cengage Learning EMEA, London, UK, 2012.

[16] T. Andersson and P. Värbrand. Decision support tools for ambulance dispatch and relocation. *Journal of the Operational Research Society*, 58(2):195–201, 2006.

[17] R.N. Anthony. *Planning and control systems: a framework for analysis*. Division of Research, Graduate School of Business Administration, Harvard University, Boston, MA, USA, 1965.

[18] J.P. Arnaout. Heuristics for the maximization of operating rooms utilization using simulation. *Simulation*, 86(8-9):573–583, 2010.

[19] M. Asaduzzaman, T.J. Chaussalet, and N.J. Robertson. A loss network model with overflow for capacity planning of a neonatal unit. *Annals of Operations Research*, 178(1):67–76, 2010.

[20] R. Ashton, L. Hague, M. Brandreth, D. Worthington, and S. Cropper. A simulation-based study of a NHS walk-in centre. *Journal of the Operational Research Society*, 56(2):153–161, 2005.

[21] V. Augusto, X. Xie, and V. Perdomo. Operating theatre scheduling with patient recovery in both operating rooms and recovery beds. *Computers & Industrial Engineering*, 58(2):231–238, 2010.

[22] M. Babes and G.V. Sarma. Out-patient queues at the Ibn-Rochd health centre. *Journal of the Operational Research Society*, 42(10):845–855, 1991.

[23] A. Bagust, M. Place, and J.W. Posnett. Dynamics of bed use in accommodating emergency admissions: stochastic simulation model. *British Medical Journal*, 319(7203):155–158, 1999.

[24] N.T.J. Bailey. On queueing processes with bulk service. *Journal of the Royal Statistical Society Series B (Methodological)*, 189(1):80–87, 1954.

[25] G. Balbo. Introduction to stochastic Petri nets. In E. Brinksma, H. Hermanns, and J.P. Katoen, editors, *Lectures on formal methods and performance analysis*, volume 2090 of *Lecture Notes in Computer Science*, pages 84–155. 2001.

[26] M.O. Ball and F.L. Lin. A reliability model applied to emergency service vehicle location. *Operations Research*, 41(1):18–36, 1993.

[27] J. Barado, J.M. Guergué, L. Esparza, C. Azcárate, F. Mallor, and S. Ochoa. A mathematical model for simulating daily bed occupancy in an intensive care unit. *Critical Care Medicine*, 40(4):1098–1104, 2011.

[28] J.F. Bard and H.W. Purnomo. Hospital-wide reactive scheduling of nurses with preference considerations. *IIE Transactions*, 37(7):589–608, 2005.

[29] E.R. Barthel, J.R. Pierce, C.J. Goodhue, H.R. Ford, T.C. Grikscheit, and J.S. Upperman. Availability of a pediatric trauma center in a disaster surge decreases triage time of the pediatric surge population: a population kinetics model. *Theoretical Biology and Medical Modelling*, 8(1):38, 2011.

[30] A. Başar, B. Çatay, and T. Ünlüyurt. A multi-period double coverage approach for locating the emergency medical service stations in Istanbul. *Journal of the Operational Research Society*, 62(4):627–637, 2010.

[31] F. Baskett, K.M. Chandy, R.R. Muntz, and F.G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *Journal of the ACM*, 22(2):248–260, 1975.

[32] S. Batun, B.T. Denton, T.R. Huschka, and A.J. Schaefer. Operating room pooling and parallel surgery processing under uncertainty. *INFORMS Journal on Computing*, 23(2):220–237, 2011.

[33] K.S. Bay, P. Leatt, and S.M. Stinson. A patient-classification system for long-term care. *Medical Care*, 20(5):468–488, 1982.

[34] P. Beattie, M. Dowda, C. Turner, L. Michener, and R. Nelson. Longitudinal continuity of care is associated with high patient satisfaction with physical therapy. *Physical Therapy*, 85(10):1046–1052, 2005.

[35] H. Beaulieu, J.A. Ferland, B. Gendron, and P. Michelon. A mathematical programming approach for scheduling physicians in the emergency room. *Health Care Management Science*, 3(3):193–200, 2000.

[36] R. Beech, R.L. Brough, and B.A. Fitzsimons. The development of a decision-support system for planning services within hospitals. *Journal of the Operational Research Society*, 41(11):995–1006, 1990.

[37] S.V. Begur, D.M. Miller, and J.R. Weaver. An integrated spatial DSS for scheduling and routing home-health-care nurses. *Interfaces*, 27(4):35–48, 1997.

[38] R. Bekker and A.M. de Bruin. Time-dependent analysis for refused admissions in clinical wards. *Annals of Operations Research*, 178(1):45–65, 2010.

[39] R. Bekker and P.M. Koeleman. Scheduling admissions and reducing variability in bed demand. *Health Care Management Science*, 14(3):1–13, 2011.

[40] J. Beliën and E. Demeulemeester. Building cyclic master surgery schedules with leveled resulting bed occupancy. *European Journal of Operational Research*, 176(2):1185 – 1204, 2007.

[41] J. Beliën and E. Demeulemeester. A branch-and-price approach for integrating nurse and surgery scheduling. *European Journal of Operational Research*, 189(3):652–668, 2008.

[42] J. Beliën, E. Demeulemeester, and B. Cardoen. A decision support system for cyclic master surgery scheduling with multiple objectives. *Journal of Scheduling*, 12(2):147–161, 2009.

[43] J.C. Bennett and D.J. Worthington. An example of a good but partially successful OR engagement: Improving outpatient clinic operations. *Interfaces*, 28(5):56–69, 1998.

[44] R. Benveniste. Solving the combined zoning and location problem for several emergency units. *Journal of the Operational Research Society*, 36(5):433–450, 1985.

[45] E. Benzarti, E. Sahin, and Y. Dallery. A literature review on operations management based models developed for home health care services. Technical report, Cahier d'Études et de Recherche, Ecole Centrale Paris, 2010.

[46] P. Beraldi and M.E. Bruni. A probabilistic model applied to emergency service vehicle location. *European Journal of Operational Research*, 196(1):323–331, 2009.

[47] P. Beraldi, M.E. Bruni, and D. Conforti. Designing robust emergency medical service via stochastic programming. *European Journal of Operational Research*, 158(1):183–193, 2004.

[48] G.N. Berlin, C. ReVelle, and D.J. Elzinga. Determining ambulance-hospital locations for on-scene and hospital services. *Environment and Planning A*, 8(5):553–561, 1976.

[49] S. Bertels and T. Fahle. A hybrid setup for a hybrid scenario: combining heuristics for the home health care problem. *Computers & Operations Research*, 33(10):2866–2890, 2006.

[50] B. Berthomieu and M. Diaz. Modeling and verification of time dependent systems using time Petri nets. *IEEE Transactions on Software Engineering*, 17(3):259–273, 1991.

[51] J.W.M. Bertrand, J.C. Wortmann, and J. Wijngaard. *Production control: a structural and design oriented approach*. Elsevier, New York, NY, USA, 1990.

[52] M.J. Bester, I. Nieuwoudt, and J.H. van Vuuren. Finding good nurse duty schedules: a case study. *Journal of Scheduling*, 10(6):387–405, 2007.

[53] G. Bianchi and R.L. Church. A hybrid fleet model for emergency medical service system design. *Social Science & Medicine*, 26(1):163–171, 1988.

[54] J. Billington, M. Diaz, and G. Rozenberg, editors. *Application of Petri nets to communication networks: advances in Petri nets*, volume 1605 of *Lecture Notes in Computer Science*. 1999.

[55] J.F. Bithell. A class of discrete-time models for the study of hospital admission systems. *Operations Research*, 17(1):48–69, 1969.

[56] E.L. Blair and C.E. Eric. A queueing network approach to health care planning with an application to burn care in New York state. *Socio-economic Planning Sciences*, 15(5):207–216, 1981.

[57] M. Blais, S.D. Lapierre, and G. Laporte. Solving a home-care districting problem in an urban setting. *Journal of the Operational Research Society*, 54(11):1141–1147, 2003.

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

[59] J.T. Blake and M.W. Carter. A goal programming approach to strategic resource allocation in acute care hospitals. *European Journal of Operational Research*, 140(3):541–561, 2002.

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

[61] J.T. Blake and J. Donald. Mount Sinai hospital uses integer programming to allocate operating room time. *Interfaces*, 32(2):63–73, 2002.

[62] D. Boldy and N. Howell. The geographical allocation of community care resources – a case study. *Journal of the Operational Research Society*, 31(2):123–129, 1980.

[63] S. Borst, A. Mandelbaum, and M.I. Reiman. Dimensioning large call centers. *Operations Research*, 52(1):17–34, 2004.

[64] R.J. Boucherie. A characterization of independence for competing Markov chains with applications to stochastic Petri nets. In *Proceedings of the 5th International Workshop on Petri Nets and Performance Models*, pages 117–126, 1993.

[65] R.J. Boucherie. Norton's equivalent for queueing networks comprised of quasi-reversible components linked by state-dependent routing. *Performance Evaluation*, 32(2):83–99, 1998.

[66] R.J. Boucherie and M. Sereno. A structural characterisation of product form stochastic Petri nets. In *Performance Evaluation of Parallel and Distributed Systems: Solution Methods: Proceedings of the 3rd QMIPS Workshop*, pages 157–174, 1994.

[67] R.J. Boucherie and M. Sereno. On the traffic equations for batch routing queueing networks and stochastic Petri nets. Technical Report 04/94-R032, CWI, Amsterdam, the Netherlands, 1994.

[68] R.J. Boucherie and M. Sereno. On closed support *T*-invariants and the traffic equations. *Journal of Applied Probability*, 35(2):473–481, 1998.

[69] R.J. Boucherie and N.M. van Dijk. Product forms for queueing networks with state-dependent multiple job transitions. *Advances in Applied Probability*, 23(1):152–187, 1991.

[70] R.J. Boucherie and N.M. van Dijk. A generalization of Norton's theorem for queueing networks. *Queueing Systems*, 13(1):251–289, 1993.

[71] J. Bowers and G. Mould. Concentration and the variability of orthopaedic demand. *Journal of the Operational Research Society*, 53(2):203–210, 2002.

[72] J. Bowers and G. Mould. Managing uncertainty in orthopaedic trauma theatres. *European Journal of Operational Research*, 154(3):599–608, 2004.

[73] M. Brahimi and D.J. Worthington. Queueing models for out-patient appointment systems – a case study. *Journal of the Operational Research Society*, 42(9):733–746, 1991.

[74] S.C. Brailsford, P.R. 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.

[75] S.C. Brailsford, V.A. Lattimer, P. Tarnaras, and J.C. Turnbull. Emergency and on-demand health care: modelling a large complex system. *Journal of the Operational Research Society*, 55(1):34–42, 2004.

[76] S.C. Brailsford and J.M.H. Vissers. OR in healthcare: a European perspective. *European Journal of Operational Research*, 212(2):223 – 234, 2011.

[77] M.L. Brandeau, F. Sainfort, and W.P. Pierskalla, editors. *Operations research and health care: a handbook of methods and applications*, volume 70 of *International Series in Operations Research & Management Science*. Kluwer Academic Publishers, Dordrecht, the Netherlands, 2004.

[78] A. Brandt. *On Norton's theorem for multi-class queueing networks of quasi reversible nodes*. Preprint 256, Sektion Mathematik, Humboldt-Universität, Berlin, Germany, 1990.

[79] A. Brandwajn. Equivalence and decomposition in queueing systems – a unified approach. *Performance Evaluation*, 5(3):175–186, 1985.

[80] O. Bräysy, W. Dullaert, and P. Nakari. The potential of optimization in communal routing problems: case studies from Finland. *Journal of Transport Geography*, 17(6):484–490, 2009.

[81] O. Bräysy, P. Nakari, W. Dullaert, and P. Neittaanmäki. An optimization approach for communal home meal delivery service: a case study. *Journal of Computational and Applied Mathematics*, 232(1):46–53, 2009.

[82] D. Bredström and M. Rönnqvist. Combined vehicle routing and scheduling with temporal precedence and synchronization constraints. *European Journal of Operational Research*, 191(1):19–31, 2008.

[83] K.M. Bretthauer, H.S. Heese, H. Pun, and E. Coe. Blocking in healthcare operations: a new heuristic and an application. *Production and Operations Management*, 20(3):375–391, 2011.

[84] L. Brotcorne, G. Laporte, and F. Semet. Ambulance location and relocation models. *European Journal of Operational Research*, 147(3):451–463, 2003.

[85] J.R. Broyles, J.K. Cochran, and D.C. Montgomery. A statistical Markov chain approximation of transient hospital inpatient inventory. *European Journal of Operational Research*, 207(3):1645–1657, 2010.

[86] H. Bruneel. Performance of discrete-time queueing systems. *Computers & Operations Research*, 20(3):303–320, 1993.

[87] H. Bruneel and I. Wuyts. Analysis of discrete-time multiserver queueing models with constant service times. *Operations Research Letters*, 15(5):231–236, 1994.

[88] J.O. Brunner, J.F. Bard, and R. Kolisch. Midterm scheduling of physicians with flexible shifts using branch and price. *IIE Transactions*, 43(2):84–109, 2011.

[89] J.O. Brunner and G. Edenharter. Long term staff scheduling of physicians with different experience levels in hospitals using column generation. *Health Care Management Science*, 14(2):189–202, 2011.

[90] J. Buchan and M.R. Dal Poz. Skill mix in the health care workforce: reviewing the evidence. *Bull World Health Organ*, 80(7):575–580, 2002.

[91] E.K. Burke, P. de Causmaecker, G.V. Berghe, and H. van Landeghem. The state of the art of nurse rostering. *Journal of Scheduling*, 7(6):441–499, 2004.

[92] C.R. Busby and M.W. Carter. A decision tool for negotiating home care funding levels in Ontario. *Home Health Care Services Quarterly*, 25(3-4):91, 2006.

[93] T.W. Butler, K.R. Karwan, J.R. Sweigart, and G.R. Reeves. An integrative model-based approach to hospital layout. *IIE transactions*, 24(2):144–152, 1992.

[94] A.B. Calvo and D.H. Marks. Location of health care facilities: an analytical approach. *Socio-Economic Planning Sciences*, 7(5):407–422, 1973.

[95] H. Campbell, R. Hotchkiss, N. Bradshaw, and M. Porteous. Integrated care pathways. *British Medical Journal*, 316(7125):133–137, 1998.

[96] B. Cardoen and E. Demeulemeester. A decision support system for surgery sequencing at UZ Leuven's day-care department. *International Journal of Information Technology and Decision Making*, 10(3):435, 2011.

[97] B. Cardoen, E. Demeulemeester, and J. Beliën. Optimizing a multiple objective surgical case sequencing problem. *International Journal of Production Economics*, 119(2):354–366, 2009.

[98] B. Cardoen, E. Demeulemeester, and J. Beliën. Sequencing surgical cases in a day-care environment: an exact branch-and-price approach. *Computers and Operations Research*, 36(9):2660–2669, 2009.

[99] 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.

[100] A.P. Carpenter, L.M. Leemis, A.S. Papir, D.J. Phillips, and G.S. Phillips. Managing magnetic resonance imaging machines: support tools for scheduling and planning. *Health Care Management Science*, 14(2):158–173, 2011.

[101] G.M. Carter, J.M. Chaiken, and E. Ignall. Response areas for two emergency units. *Operations Research*, 20(3):571–594, 1972.

[102] M.W. Carter. Diagnosis: mismanagement of resources. *OR/MS Today*, 29(2):26–33, 2002.

[103] M.W. Carter and S.D. Lapierre. Scheduling emergency room physicians. *Health Care Management Science*, 4(4):347–360, 2001.

[104] T. Cayirli and E. Veral. Outpatient scheduling in health care: a review of literature. *Production and Operations Management*, 12(4):519–549, 2003.

[105] T. Cayirli, E. Veral, and Rosen H. Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science*, 9(1):47–58, 2006.

[106] T. Cayirli, E. Veral, and H. Rosen. Assessment of patient classification in appointment system design. *Production and Operations Management*, 17(3):338–353, 2008.

[107] CC Zorgadviseurs. Revalidatie in beweging [Rehabilitation on the move]. *Retrieved October 13, 2012, from http://www.revalidatienederland.nl/actueel/publicaties*, 2010.

[108] R. Ceglowski, L. Churilov, and J. Wasserthiel. Combining data mining and discrete event simulation for a value-added view of a hospital emergency department. *Journal of the Operational Research Society*, 58(2):246–254, 2006.

[109] Centraal Planbureau [Netherlands Bureau for Economic Policy Analysis]. Trends in gezondheid en zorg [Trends in health and healthcare]. *Retreived, October 13, 2012, from http://www.cpb.nl/publicatie/trends-in-gezondheid-en-zorg*, 2011.

[110] S. Ceschia and A. Schaerf. Local search and lower bounds for the patient admission scheduling problem. *Computers and Operations Research*, 38(10):1452–1463, 2011.

[111] S. Chaabane, N. Meskens, A. Guinet, and M. Laurent. Comparison of two methods of operating theatre planning: application in Belgian hospital. *Journal of Systems Science and Systems Engineering*, 17(2):171–186, 2008.

[112] S. Chahed, E. Marcon, E. Sahin, D. Feillet, and Y. Dallery. Exploring new operational research opportunities within the home care context: the chemotherapy at home. *Health Care Management Science*, 12(2):179–191, 2009.

[113] S. Chand, H. Moskowitz, J.B. Norris, S. Shade, and D.R. Willis. Improving patient flow at an outpatient clinic: study of sources of variability and improvement factors. *Health Care Management Science*, 12(3):325–340, 2009.

[114] K.M. Chandy, U. Herzog, and L. Woo. Parametric analysis of queuing networks. *IBM Journal of Research and Development*, 19(1):36–42, 2010.

[115] X. Chao. Networks with customers, signals and product form solutions. In R.J. Boucherie and N.M. Van Dijk, editors, *Queueing networks: a fundamental approach*, pages 217–268. Springer, New York, NY, 2010.

[116] M.L. Chaudhry, B.R. Madill, and G. Briere. Computational analysis of steady-state probabilities of $M|G^{a,b}|1$ and related nonbulk queues. *Queueing Systems*, 2(2):93–114, 1987.

[117] T.J. Chaussalet, H. Xie, and P.H. Millard. A closed queueing network approach to the analysis of patient flow in health care systems. *Methods of Information in Medicine*, 45(5):492–497, 2006.

[118] B. Cheang, H. Li, A. Lim, and B. Rodrigues. Nurse rostering problems – a bibliographic survey. *European Journal of Operational Research*, 151(3):447–460, 2003.

[119] C.F. Chien, Y.C. Huang, and C.H. 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, 2009.

[120] C.F. Chien, F.P. Tseng, and C.H. Chen. An evolutionary approach to rehabilitation patient scheduling: a case study. *European Journal of Operational Research*, 189(3):1234–1253, 2008.

[121] A. Chockalingam, K. Jayakumar, and M.A. Lawley. A stochastic control approach to avoiding emergency department overcrowding. In *Proceedings of the Winter Simulation Conference*, pages 2399–2411, 2010.

[122] V.S. Chow, M.L. Puterman, N. Salehirad, W. Huang, and D. Atkins. Reducing surgical ward congestion through improved surgical scheduling and uncapacitated simulation. *Production and Operations Management*, 20(3):418–430, 2011.

[123] G. Christodoulou and G.J. Taylor. Using a continuous time hidden Markov process, with covariates, to model bed occupancy of people aged over 65 years. *Health Care Management Science*, 4(1):21–24, 2001.

[124] J.K. Cochran and A. Bharti. Stochastic bed balancing of an obstetrics hospital. *Health Care Management Science*, 9(1):31–45, 2006.

[125] J.K. Cochran and K. Roche. A queuing-based decision support methodology to estimate hospital inpatient bed demand. *Journal of the Operational Research Society*, 59(11):1471–1482, 2007.

[126] J.K. Cochran and K.T. Roche. A multi-class queuing network analysis methodology for improving hospital emergency department performance. *Computers & Operations Research*, 36(5):1497–1512, 2009.

[127] T. Coelli, D.S. Prasada Rao, and G.E. Battese. *An introduction to efficiency and productivity analysis*. Springer, New York, NY, USA, 2005.

[128] M.A. Cohen, J.C. Hershey, and E.N. Weiss. Analysis of capacity decisions for progressive patient care hospital facilities. *Health Services Research*, 15(2):145–160, 1980.

[129] J.L. Coleman. *Stochastic Petri nets with product form equilibrium distributions*. PhD thesis, University of Adelaide, Australia, 1993.

[130] J.L. Coleman, W. Henderson, and P.G. Taylor. Product form equilibrium distributions and a convolution algorithm for stochastic Petri nets. *Performance Evaluation*, 26(3):159–180, 1996.

[131] Commission on Accreditation of Rehabilitation Facilities (CARF) International. Homepage. *Retrieved October 13, 2012, from: http://www.carf.org/home/*.

[132] D. Conforti, F. Guerriero, and R. Guido. Optimization models for radiotherapy patient scheduling. *4OR: A Quarterly Journal of Operations Research*, 6(3):263–278, 2008.

[133] D. Conforti, F. Guerriero, and R. Guido. Non-block scheduling with priority for radiotherapy treatments. *European Journal of Operational Research*, 201(1):289–296, 2010.

[134] D. Conforti, F. Guerriero, R. Guido, M.M. Cerinic, and M.L. Conforti. An optimal decision making model for supporting week hospital management. *Health Care Management Science*, 14(1):74–88, 2011.

[135] D. Conforti, F. Guerriero, R. Guido, and M. Veltri. An optimal decision-making approach for the management of radiotherapy patients. *OR Spectrum*, 33(1):123–148, 2011.

[136] A.X. Costa, S.A. Ridley, A.K. Shahani, P.R. Harper, V. De Senna, and M.S. Nielsen. Mathematical modelling and simulation for planning critical care capacity. *Anaesthesia*, 58(4):320–327, 2003.

[137] M.J. Côté, S.S. Syam, W.B. Vogel, and D.C. Cowper. A mixed integer programming model to locate traumatic brain injury treatment units in the department of veterans affairs: a case study. *Health Care Management Science*, 10(3):253–267, 2007.

[138] S. Creemers. *Appointment-driven queueing systems*. PhD thesis, Katholieke Universiteit Leuven, Belgium, 2009.

[139] S. Creemers and M. Lambrecht. An advanced queueing model to analyze appointment-driven service systems. *Computers & Operations Research*, 36(10):2773–2785, 2009.

[140] S. Creemers and M. Lambrecht. Queueing models for appointment-driven systems. *Annals of Operations Research*, 178(1):155–172, 2010.

[141] M. Criswell, I. Hasan, R. Kopach, S. Lambert, M. Lawley, D. McWilliams, G. Trupiano, and N. Varadarajan. Emergency department divert avoidance using Petri nets. In *IEEE International Conference on System of Systems Engineering*, pages 1–6, 2007.

[142] V. Currie, G. Harvey, E. West, H. McKenna, and S. Keeney. Relationship between quality of care, staffing levels, skill mix and nurse autonomy: literature review. *Journal of Advanced Nursing*, 51(1):73–82, 2005.

[143] H. Daduna. Discrete time queueing networks with product form steady state. In R.J. Boucherie and N.M. van Dijk, editors, *Queueing networks: a fundamental approach*, pages 269–312. Springer, New York, NY, USA, 2010.

[144] W. Dafoe, H. Arthur, H. Stokes, L. Morrin, and L. Beaton. Universal access: but when? treating the right patient at the right time: access to cardiac rehabilitation. *Canadian Journal of Cardiology*, 22(11):905–911, 2006.

[145] M.S. Daskin and E.H. Stern. A hierarchical objective set covering model for emergency medical service vehicle deployment. *Transportation Science*, 15(2):137, 1981.

[146] R.W. Day, M.D. 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.

[147] V. De Angelis. Planning home assistance for AIDS patients in the city of Rome, Italy. *Interfaces*, 48(3):75–83, 1998.

[148] L. de Bleser, R. Depreitere, K.D.E. Waele, K. Vanhaecht, J. Vlayen, and W. Sermeus. Defining pathways. *Journal of Nursing Management*, 14(7):553–563, 2006.

[149] A.M. de Bruin, A.C. van Rossum, M.C. Visser, and G.M. Koole. Modeling the emergency cardiac in-patient flow: an application of queuing theory. *Health Care Management Science*, 10(2):125–137, 2007.

[150] P. de Causmaecker and G. vanden Berghe. A categorisation of nurse rostering problems. *Journal of Scheduling*, 14(1):3–16, 2011.

[151] M.L. De Grano, D.J. Medeiros, and D. Eitel. Accommodating individual preferences in nurse scheduling via auctions and optimization. *Health Care Management Science*, 12(3):228–242, 2009.

[152] N. Dellaert, J. Jeunet, and G. Mincsovics. Budget allocation for permanent and contingent capacity under stochastic demand. *International Journal of Production Economics*, 131(1):128–138, 2011.

[153] 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.

[154] B.T. Denton, O. Alagoz, A. Holder, and E.K. Lee. Medical decision making: open research challenges. *IIE Transactions on Healthcare Systems Engineering*, 1(3):161–167, 2011.

[155] B.T. Denton and D. Gupta. A sequential bounding approach for optimal appointment scheduling. *IIE Transactions*, 35(11):1003–1016, 2003.

[156] B.T. Denton, A.J. Miller, H.J. Balasubramanian, and T.R. Huschka. Optimal allocation of surgery blocks to operating rooms under uncertainty. *Operations research*, 58(4-Part-1):802–816, 2010.

[157] B.T. Denton, J. Viapiano, and A. Vogl. Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Science*, 10(1):13–24, 2007.

[158] M.S. Desai, M.L. Penn, S.C. Brailsford, and M. Chipulu. Modelling of Hampshire adult services – gearing up for future demands. *Health Care Management Science*, 11(2):167–176, 2008.

[159] A.A. Desrochers and R.Y. Al-Jaar. *Applications of Petri nets in manufacturing systems: modeling, control, and performance analysis*. IEEE Press, New York, NY, USA, 1995.

[160] F. Dexter. Design of appointment systems for preanesthesia evaluation clinics to minimize patient waiting times: a review of computer simulation and patient survey studies. *Anesthesia & Analgesia*, 89(4):925–931, 1999.

[161] F. Dexter. Bibliography of operating room management articles. *Retrieved October 13, 2012, from http://www.franklindexter.com/*, 2012.

[162] F. Dexter, R.H. Epstein, and H.M. Marsh. A statistical analysis of weekday operating room anesthesia group staffing costs at nine independently managed surgical suites. *Anesthesia & Analgesia*, 92(6):1493–1498, 2001.

[163] F. Dexter and J. Ledolter. Bayesian prediction bounds and comparisons of operating room times even for procedures with few or no historic data. *Anesthesiology*, 103(6):1259–1267, 2005.

[164] F. Dexter and A. Macario. Decrease in case duration required to complete an additional case during regularly scheduled hours in an operating room suite: a computer simulation study. *Anesthesia & Analgesia*, 88(1):72–76, 1999.

[165] F. Dexter, A. Macario, and D.A. Lubarsky. The impact on revenue of increasing patient volume at surgical suites with relatively high operating room utilization. *Anesthesia & Analgesia*, 92(5):1215–1221, 2001.

[166] F. Dexter, A. Macario, and L. O'Neill. A strategy for deciding operating room assignments for second-shift anesthetists. *Anesthesia & Analgesia*, 89(4):920–924, 1999.

[167] F. Dexter, A. Macario, and L. O'Neill. Scheduling surgical cases into overflow block time – computer simulation of the effects of scheduling strategies on operating room labor costs. *Anesthesia & Analgesia*, 90(4):980–988, 2000.

[168] F. Dexter, A. Macario, and R.D. Traub. Optimal sequencing of urgent surgical cases. *Journal of Clinical Monitoring and Computing*, 15(3):153–162, 1999.

[169] F. Dexter, A. Macario, and R.D. Traub. Which algorithm for scheduling add-on elective cases maximizes operating room utilization?: Use of bin packing algorithms and fuzzy constraints in operating room management. *Anesthesiology*, 91(5):1491–1500, 1999.

[170] F. Dexter, A. Macario, R.D. Traub, M. Hopwood, and D.A. Lubarsky. An operating room scheduling strategy to maximize the use of operating room block time: computer simulation of patient scheduling and survey of patients' preferences for surgical waiting time. *Anesthesia & Analgesia*, 89(1):7–20, 1999.

[171] F. Dexter, A. Macario, R.D. Traub, and D.A. Lubarsky. Operating room utilization alone is not an accurate metric for the allocation of operating room block time to individual surgeons with low caseloads. *Anesthesiology*, 98(5):1243–1249, 2003.

[172] F. Dexter and R.D. Traub. Statistical method for predicting when patients should be ready on the day of surgery. *Anesthesiology*, 93(4):1107, 2000.

[173] F. Dexter and R.D. Traub. How to schedule elective surgical cases into specific operating rooms to maximize the efficiency of use of operating room time. *Anesthesia & Analgesia*, 94(4):933–942, 2002.

[174] F. Dexter, R.D. Traub, and P. Lebowitz. Scheduling a delay between different surgeons' cases in the same operating room on the same day using upper prediction bounds for case durations. *Anesthesia & Analgesia*, 92(4):943–946, 2001.

[175] F. Dexter, R.D. Traub, and A. Macario. How to release allocated operating room time to increase efficiency: predicting which surgical service will have the most underutilized operating room time. *Anesthesia & Analgesia*, 96(2):507–512, 2003.

[176] F. Dexter, R.E. Wachtel, R.H. Epstein, J. Ledolter, and M.M. Todd. Analysis of operating room allocations to optimize scheduling of specialty rotations for anesthesia trainees. *Anesthesia & Analgesia*, 111(2):520–524, 2010.

[177] F. Dexter, R.E. Wachtel, M.W. Sohn, J. Ledolter, E.U. Dexter, and A. Macario. Quantifying effect of a hospital's caseload for a surgical specialty on that of another hospital using multi-attribute market segments. *Health Care Management Science*, 8(2):121–131, 2005.

[178] M. Diaz and P. Azema. Petri net based models for the specification and validation of protocols. In G. Goos and J. Hartmanis, editors, *Advances in Petri nets*, volume 188 of *Lecture Notes in Computer Science*, pages 101–121. 1985.

[179] G. Dobson, S. Hasija, and E.J. Pinker. Reserving capacity for urgent patients in primary care. *Production and Operations Management*, 20(3):456–473, 2011.

[180] G. Dobson, H.H. Lee, and E. Pinker. A model of ICU bumping. *Operations Research*, 58(6):1564–1576, 2010.

[181] V.F. Dokmeci. Planning ambulatory health care delivery systems. *Omega*, 4(5):617–622, 1976.

[182] S. Donatelli and M. Sereno. On the product form solution for stochastic Petri nets. In K. Jensen, editor, *Proceedings of the 13th international conference on application and theory of Petri nets*, volume 616 of *Lecture Notes in Computer Science*, pages 154–172. 1992.

[183] M. Dong and F.F. Chen. Process modeling and analysis of manufacturing supply chain networks using object-oriented Petri nets. *Robotics and Computer-Integrated Manufacturing*, 17(1):121–129, 2001.

[184] K. Dong-Guen and K. Yeong-Dae. A branch and bound algorithm for determining locations of long-term care facilities. *European Journal of Operational Research*, 206(1):168–177, 2010.

[185] M.B. Dumas. Simulation modeling for hospital bed planning. *Simulation*, 43(2):69, 1984.

[186] M.B. Dumas. Hospital bed utilization: an implemented simulation approach to adjusting and maintaining appropriate levels. *Health Services Research*, 20(1):43, 1985.

[187] D.J. Eaton. Determining ambulance deployment in Santo Domingo, Dominican Republic. *Journal of the Operational Research Society*, 37(2):113–126, 1986.

[188] EBSCO Publishing. Business source elite. *Retrieved October 13, 2012, from http://www.ebscohost.com/academic/business-source-elite*.

[189] G.M. Edward, S.F. Das, S.G. Elkhuizen, P.J.M. Bakker, J.A.M. Hontelez, M.W. Hollmann, B. Preckel, and L.C. Lemaire. Simulation to analyse planning difficulties at the preoperative assessment clinic. *British Journal of Anaesthesia*, 100(2):195–202, 2008.

[190] E. El-Darzi, C. Vasilakis, T.J. Chaussalet, and P.H. Millard. A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. *Health Care Management Science*, 1(2):143–149, 1998.

[191] S.G. Elkhuizen. *Patient oriented logistics: studies on organizational improvement in an academic hospital*. PhD thesis, University of Amsterdam, the Netherlands, 2007.

[192] S.G. Elkhuizen, G. Bor, M. Smeenk, N.S. Klazinga, and P.J.M. Bakker. Capacity management of nursing staff as a vehicle for organizational improvement. *BMC health services research*, 7(1):196–205, 2007.

[193] S.G. Elkhuizen, S.F. Das, P.J.M. Bakker, and J.A.M. Hontelez. Using computer simulation to reduce access time for outpatient departments. *British Medical Journal*, 16(5):382–386, 2007.

[194] Elsevier. Scopus database. *Retrieved October 13, 2012, from http://www.scopus.com/*.

[195] G. Erdogan, E. Erkut, A. Ingolfsson, and G. Laporte. Scheduling ambulance crews for maximum coverage. *Journal of the Operational Research Society*, 61(4):543–550, 2010.

[196] E. Erkut, A. Ingolfsson, and G. Erdoğan. Ambulance location for maximum survival. *Naval Research Logistics*, 55(1):42–58, 2008.

[197] A.T. Ernst, H. Jiang, M. Krishnamoorthy, and D. Sier. Staff scheduling and rostering: a review of applications, methods and models. *European Journal of Operational Research*, 153(1):3–27, 2004.

[198] A.O. Esogbue and A.J. Singh. A stochastic model for an optimal priority bed distribution problem in a hospital ward. *Operations Research*, 24(5):884–898, 1976.

[199] J. Esparza. Decidability and complexity of Petri net problems: an introduction. In Rozenberg, editor, *Advances in Petri nets*, volume 1491 of *Lecture Notes in Computer Science*, pages 374–428. 1998.

[200] European Neuromuscular Centre. Homepage. *Retrieved October 13, 2012, from http://www.enmc.org*.

[201] P. Eveborn, P. Flisberg, and M. Ronnqvist. Laps care – an operational system for staff planning of home care. *European Journal of Operational Research*, 171(3):962–976, 2006.

[202] P. Eveborn, M. Rönnqvist, H. Einarsdóttir, M. Eklund, K. Lidén, and M. Almroth. Operations research improves quality and efficiency in home care. *Interfaces*, 39(1):18–34, 2009.

[203] J.E. Everett. A decision support simulation model for the management of an elective surgery waiting system. *Health Care Management Science*, 5(2):89–95, 2002.

[204] J. Ezpeleta, J.M. Colom, and J. Martinez. A Petri net based deadlock prevention policy for flexible manufacturing systems. *IEEE Transactions on Robotics and Automation*, 11(2):173–184, 1995.

[205] M.J. Faddy, N. Graves, and A. Pettitt. Modeling length of stay in hospital and other right skewed data: comparison of phase-type, gamma and log-normal distributions. *Value in Health*, 12(2):309–314, 2009.

[206] M.J. Faddy and S.I. McClean. Markov chain modelling for geriatric patient care. *Methods of Information in Medicine-Methodik der Information in der Medizin*, 44(3):369–373, 2005.

[207] M.J. Faddy and S.I. McClean. Using a multi-state model to enhance understanding of geriatric patient care. *Australian Health Review*, 31(1):91–97, 2007.

[208] H. Fei, C. Chu, and N. Meskens. Solving a tactical operating room planning problem by a column-generation-based heuristic procedure with four criteria. *Annals of Operations Research*, 166(1):91–108, 2009.

[209] H. Fei, C. Chu, N. Meskens, and A. Artiba. Solving surgical cases assignment problem by a branch-and-price approach. *International Journal of Production Economics*, 112(1):96–108, 2008.

[210] H. Fei, N. Meskens, and C. Chu. A planning and scheduling problem for an operating theatre using an open scheduling strategy. *Computers & Industrial Engineering*, 58(2):221–230, 2010.

[211] M. Feinberg. Chemical reaction network structure and the stability of complex isothermal reactors—1. The deficiency zero and deficiency one theorems. *Chemical Engineering Science*, 42(10):2229–2268, 1987.

[212] R.B. Fetter and J.D. Thompson. The simulation of hospital systems. *Operations Research*, 13(5):689–711, 1965.

[213] R.B. Fetter and J.D. Thompson. Patients' waiting time and doctors' idle time in the outpatient setting. *Health Services Research*, 1(1):66–90, 1966.

[214] J.A. Fitzsimmons. A methodology for emergency ambulance deployment. *Management Science*, 19(6):627–636, 1973.

[215] A. Fletcher, D. Halsall, S. Huxham, and D. Worthington. The DH accident and emergency department model: a national generic model used locally. *Journal of the Operational Research Society*, 58(12):1554–1562, 2006.

[216] G. Florin and S. Natkin. Matrix product form solution for closed synchronized queuing networks. In *The Proceedings of the 3rd International Workshop on Petri Nets and Performance Models*, pages 29–37, 1989.

[217] G. Florin and S. Natkin. Necessary and sufficient ergodicity condition for open synchronized queueing networks. *IEEE Transactions on Software Engineering*, 15(4):367, 1989.

[218] 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, 2003.

[219] B.E. Fries. Bibliography of operations research in health-care systems. *Operations Research*, 25(5):801–814, 1976.

[220] B.E. Fries and V.P. Marathe. Determination of optimal variable-sized multiple-block appointment systems. *Operations Research*, 29(2):324–345, 1981.

[221] B.E. Fries, D.P. Schneider, W.J. Foley, M. Gavazzi, R. Burke, and E. Cornelius. Refining a case-mix measure for nursing homes: Resource Utilization Groups (RUG-III). *Medical Care*, 32(7):668, 1994.

[222] D. Frosch. Product form solutions for closed synchronized systems of stochastic sequential processes. Technical report, Forschungsbericht Mathematik/Informatik 92-13, Universität Trier, Germany, 1992.

[223] D. Frosch and K. Natarajan. Product form solutions for closed synchronized systems of stochastic sequential processes. In *Proceedings of 1992 International Computer Symposium*, pages 392–402, 1992.

[224] O. Fujiwara, T. Makjamroen, and K.K. Gupta. Ambulance deployment analysis: a case study of Bangkok. *European Journal of Operational Research*, 31(1):9–18, 1987.

[225] P.H.P. Fung Kon Jin, M.G.W. Dijkgraaf, C.L. Alons, C. van Kuijk, L.F.M. Beenen, G.M. Koole, and J.C. Goslings. Improving CT scan capabilities with a new trauma workflow concept: simulation of hospital logistics using different CT scanner scenarios. *European Journal of Radiology*, 80(2):504–509, 2011.

[226] S. Gallivan and M. Utley. A technical note concerning emergency bed demand. *Health Care Management Science*, 14(3):1–3, 2011.

[227] S. Gallivan, M. Utley, T. Treasure, and O. Valencia. Booked inpatient admissions and hospital capacity: mathematical modelling study. *British Medical Journal*, 324(7332):280–282, 2002.

[228] S. Ganguli, J.C. Tham, and B.M.J. d'Othee. Establishing an outpatient clinic for minimally invasive vein care. *American Journal of Roentgenology*, 188(6):1506–1511, 2007.

[229] J.A. Garcia-Navarro and W.A. Thompson. Analysis of bed usage and occupancy following the introduction of geriatric rehabilitative care in a hospital in Huesca, Spain. *Health Care Management Science*, 4(1):63–66, 2001.

[230] L. Garg, S. McClean, B. Meenan, and P. Millard. A non-homogeneous discrete time Markov model for admission scheduling and resource planning in a cost or capacity constrained healthcare system. *Health Care Management Science*, 13(2):155–169, 2010.

[231] M. Gendreau, G. Laporte, and F. Semet. The maximal expected coverage relocation problem for emergency vehicles. *Journal of the Operational Research Society*, 57(1):22–28, 2005.

[232] N. Geng, X. Xie, V. Augusto, and Z. Jiang. A Monte Carlo optimization and dynamic programming approach for managing MRI examinations of stroke patients. *IEEE Transactions on Automatic Control*, 56(11):2515–2529, 2011.

[233] Y. Gerchak, D. Gupta, and M. Henig. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Science*, 42(3):321–334, 1996.

[234] N. Geroliminis, K. Kepaptsoglou, and M.G. Karlaftis. A hybrid hypercube-genetic algorithm approach for deploying many emergency response mobile units in an urban network. *European Journal of Operational Research*, 210(2):287–300, 2011.

[235] J. Gillespie, S. McClean, B. Scotney, L. Garg, M. Barton, and K. Fullerton. Costing hospital resources for stroke patients using phase-type models. *Health Care Management Science*, 14(13):1–13, 2011.

[236] A. Gnanlet and W.G. Gilland. Sequential and simultaneous decision making for optimizing health care resource flexibilities. *Decision Sciences*, 40(2):295–326, 2009.

[237] Y. Gocgun, B.W. Bresnahan, A. Ghate, and M.L. Gunn. A Markov decision process approach to multi-category patient scheduling in a diagnostic facility. *Artificial Intelligence in Medicine*, 53(2):73–81, 2011.

[238] J. Goddard and M. Tavakoli. Efficiency and welfare implications of managed public sector hospital waiting lists. *European Journal of Operational Research*, 184(2):778–792, 2008.

[239] H. Gold and P. Tran-Gia. Performance analysis of a batch service queue arising out of manufacturing system modelling. *Queueing Systems*, 14(3):413–426, 1993.

[240] J. Goldberg, R. Dietrich, J.M. Chen, M. Mitwasi, T. Valenzuela, and E. Criss. A simulation model for evaluating a set of emergency vehicle base locations: development, validation, and usage. *Socio-Economic Planning Sciences*, 24(2):125–141, 1990.

[241] J. Goldberg, R. Dietrich, J.M. Chen, M.G. Mitwasi, T. Valenzuela, and E. Criss. Validating and applying a model for locating emergency medical vehicles in Tuczon, AZ. *European Journal of Operational Research*, 49(3):308–324, 1990.

[242] J. Goldman, H.A. Knappenberger, and J.C. Eller. Evaluating bed allocation policy with computer simulation. *Health Services Research*, 3(2):119–129, 1968.

[243] W.J. Gordon and G.F. Newell. Closed queuing systems with exponential servers. *Operations Research*, 15(2):254–265, 1967.

[244] N. Görmez, M. Köksalan, and F.S. Salman. Locating disaster response facilities in Istanbul. *Journal of the Operational Research Society*, 62(7):1239–1252, 2010.

[245] F. Gorunescu, S.I. McClean, and P.H. Millard. A queueing model for bed-occupancy management and planning of hospitals. *Journal of the Operational Research Society*, 53(1):19–24, 2002.

[246] F. Gorunescu, S.I. McClean, and P.H. Millard. Using a queueing model to help plan bed allocation in a department of geriatric medicine. *Health Care Management Science*, 5(4):307–312, 2002.

[247] L. Goulding, J. Adamson, I. Watt, and J. Wright. Patient safety in patients who occupy beds on clinically inappropriate wards: a qualitative interview study with NHS staff. *BMJ Quality & Safety*, 21(3):218–224, 2012.

[248] D. Gove, D. Hewett, and A. Shahani. Towards a model for hospital case-load decision support. *Mathematical Medicine and Biology*, 12(3–4):329–338, 1995.

[249] S.C. Graves, H.S. Leff, J. Natkins, and M. Senger. A simple stochastic model for facility planning in a mental health care system. *Interfaces*, 13(5):101–110, 1983.

[250] M. Gray. Value: operations research and the new health care paradigm. *Operations Research for Health Care*, 1(1):20–21, 2012.

[251] L.V. Green. Queueing analysis in healthcare. In R.W. Hall, editor, *Patient flow: reducing delay in healthcare delivery*, volume 91 of *International Series in Operations Research & Management Science*, pages 281–307. Springer, New York, NY, USA, 2006.

[252] L.V. Green and P.J. Kolesar. Improving emergency responsiveness with management science. *Management Science*, 50(8):1001–1014, 2004.

[253] L.V. Green, P.J. Kolesar, and J. Soares. Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research*, 49(4):549–564, 2001.

[254] L.V. Green, P.J. Kolesar, and W. Whitt. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16(1):13–39, 2007.

[255] L.V. Green and V. Nguyen. Strategies for cutting hospital beds: the impact on patient service. *Health Services Research*, 36(2):421–442, 2001.

[256] L.V. Green and S. Savin. Reducing delays for medical appointments: A queueing approach. *Operations Research*, 56(6):1526–1538, 2008.

[257] L.V. Green, S. Savin, and B. Wang. Managing patient service in a diagnostic medical facility. *Operations Research*, 54(1):11–25, 2006.

[258] L.V. Green and J. Soares. Computing time-dependent waiting time probabilities in $M(t)|M|s(t)$ queueing systems. *Manufacturing & Service Operations Management*, 9(1):54–61, 2007.

[259] L.V. Green, J. Soares, J.F. Giglio, and R.A. Green. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine*, 13(1):61–68, 2006.

[260] J. Griffin, S. Xia, S. Peng, and P. Keskinocak. Improving patient flow in an obstetric unit. *Health Care Management Science*, 15(1):1–14, 2012.

[261] J.D. Griffiths, N. Price-Lloyd, M. Smithies, and J.E. Williams. Modelling the requirement for supplementary nurses in an intensive care unit. *Journal of the Operational Research Society*, 56(2):126–133, 2005.

[262] F. Guerriero and R. Guido. Operational research in the management of the operating theatre: a survey. *Health Care Management Science*, 14(1):89–114, 2011.

[263] A. Guinet and S. Chaabane. Operating theatre planning. *International Journal of Production Economics*, 85(1):69–81, 2003.

[264] S. Gul, B.T. Denton, J.W. Fowler, and T. Huschka. Bi-criteria scheduling of surgical services for an outpatient procedure center. *Production and Operations Management*, 20(3):406–417, 2011.

[265] E.D. Güneş. Modeling time allocation for prevention in primary care. *Central European Journal of Operations Research*, 17(3):359–380, 2009.

[266] D. Gupta. Surgical suites' operations management. *Production and Operations Management*, 16(6):689–700, 2007.

[267] D. Gupta and B. Denton. Appointment scheduling in health care: Challenges and opportunities. *IIE transactions*, 40(9):800–819, 2008.

[268] D. Gupta and L. Wang. Revenue management for a primary-care clinic in the presence of patient choice. *Operations Research*, 56(3):576–592, 2008.

[269] S. Haddad, J. Mairesse, and H.T. Nguyen. Synthesis and analysis of product-form Petri nets. In L.M. Kristensen and L. Petrucci, editors, *Applications and theory of Petri nets*, volume 6709 of *Lecture Notes in Computer Science*, pages 288–307. 2011.

[270] S. Haddad, P. Moreaux, M. Sereno, and M. Silva. Product-form and stochastic Petri nets: a structural approach. *Performance evaluation*, 59(4):313–336, 2005.

[271] R.W. Hall, editor. *Patient flow: reducing delay in healthcare delivery*, volume 91 of *International Series in Operations Research & Management Science*. Springer, New York, NY, USA, 2006.

[272] R.W. Hall, editor. *Handbook of healthcare system scheduling*, volume 168 of *International Series in Operations Research & Management Science*. Springer, New York, NY, USA, 2011.

[273] E.W. Hans, M. van Houdenhoven, and P.J.H. Hulshof. A framework for health care planning and control. In R.W. Hall, editor, *Handbook of healthcare system scheduling*, volume 168 of *International Series in Operations Research & Management Science*, pages 303–320. Springer, New York, NY, USA, 2012.

[274] E.W. Hans, G. Wullink, M. van Houdenhoven, and G. Kazemier. Robust surgery loading. *European Journal of Operational Research*, 185(3):1038–1050, 2008.

[275] W.L. Hare, A. Alimadad, H. Dodd, R. Ferguson, and A. Rutherford. A deterministic model of home and community care client counts in British Columbia. *Health Care Management Science*, 12(1):80–98, 2009.

[276] S.I. Harewood. Emergency ambulance deployment in Barbados: a multi-objective approach. *Journal of the Operational Research Society*, 53(2):185–192, 2002.

[277] P.R. Harper. A framework for operational modelling of hospital resources. *Health Care Management Science*, 5(3):165–173, 2002.

[278] P.R. Harper and H.M. Gamlin. Reduced outpatient waiting times with improved appointment scheduling: a simulation modelling approach. *OR Spectrum*, 25(2):207–222, 2003.

[279] P.R. Harper, V.A. Knight, and A.H. Marshall. Discrete conditional phase-type models utilising classification trees: Application to modelling health service capacities. *European Journal of Operational Research*, 219(3):522–530, 2011.

[280] P.R. Harper, N.H. Powell, and J.E. Williams. Modelling the size and skill-mix of hospital nursing teams. *Journal of the Operational Research Society*, 61(5):768–779, 2010.

[281] P.R. Harper and A.K. Shahani. Modelling for the planning and management of bed capacities in hospitals. *Journal of the Operational Research Society*, 53(1):11–18, 2002.

[282] P.R. Harper, A.K. Shahani, J.E. Gallagher, and C. Bowie. Planning health services with explicit geographical considerations: a stochastic location-allocation approach. *Omega*, 33(2):141–152, 2005.

[283] R.A. Harris. Hospital bed requirements planning. *European Journal of Operational Research*, 25(1):121–126, 1986.

[284] G.W. Harrison and G.J. Escobar. Length of stay and imminent discharge probability distributions from multistage models: variation by diagnosis, severity of illness, and hospital. *Health Care Management Science*, 13(3):268–279, 2010.

[285] G.W. Harrison and P.H. Millard. Balancing acute and long-term care: the mathematics of throughput in departments of geriatric medicine. *Methods of Information in Medicine*, 30(3):221, 1991.

[286] G.W. Harrison, A. Shafer, and M. Mackay. Modelling variability in hospital bed occupancy. *Health Care Management Science*, 8(4):325–334, 2005.

[287] P.G. Harrison. Turning back time in Markovian process algebra. *Theoretical computer science*, 290(3):1947–1986, 2003.

[288] P.G. Harrison. Compositional reversed Markov processes, with applications to G-networks. *Performance Evaluation*, 57(3):379–408, 2004.

[289] P.G. Harrison. Reversed processes, product forms and a non-product form. *Linear Algebra and Its Applications*, 386:359–381, 2004.

[290] P.G. Harrison and J. Hillston. Exploiting quasi-reversible structures in Markovian process algebra models. *The Computer Journal*, 38(7):510, 1995.

[291] R. Hassin and S. Mendel. Scheduling arrivals to queues: a single-server model with no-shows. *Management Science*, 54(3):565–572, 2008.

[292] J.E. Helm, S. AhmadBeygi, and M.P. Van Oyen. Design and analysis of hospital admission control for operational effectiveness. *Production and Operations Management*, 20(3):359–374, 2011.

[293] W. Henderson. Finding and using exact equilibrium distributions for stochastic Petri nets. *Computer Networks and ISDN Systems*, 25(10):1143–1153, 1993.

[294] W. Henderson and D. Lucic. Exact results in the aggregation and disaggregation of stochastic petri nets. In *Proceedings of the 4th International Workshop on Petri Nets and Performance Models*, pages 166–175, 1991.

[295] W. Henderson and D. Lucic. Aggregation and disaggregation through insensitivity in stochastic petri nets. *Performance Evaluation*, 17(2):91–114, 1993.

[296] W. Henderson, D. Lucic, and P.G. Taylor. A net level performance analysis of stochastic Petri nets. *The ANZIAM Journal*, 31(2):176–187, 1989.

[297] W. Henderson, C.E.M. Pearce, P.G. Taylor, and N.M. van Dijk. Closed queueing networks with batch services. *Queueing Systems*, 6(1):59–70, 1990.

[298] W. Henderson and P.G. Taylor. Aggregation methods in exact performance analysis of stochastic Petri nets. In *Proceedings of the 3rd International Workshop on Petri Nets and Performance Models*, pages 12–18, 1989.

[299] W. Henderson and P.G. Taylor. Product form in networks of queues with batch arrivals and batch services. *Queueing Systems*, 6(1):71–87, 1990.

[300] W. Henderson and P.G. Taylor. Embedded processes in stochastic Petri nets. *IEEE Transactions on Software Engineering*, 17(2):108–116, 1991.

[301] W. Herring and J. Herrmann. The single-day surgery scheduling problem: sequential decision-making and threshold-based heuristics. *OR Spectrum*, 34(2):429–459, 2012.

[302] J.C. Hershey, E.N. Weiss, and M.A. Cohen. A stochastic service network model with application to hospital facilities. *Operations Research*, 29(1):1–22, 1981.

[303] A. Hertz and N. Lahrichi. A patient assignment algorithm for home care services. *Journal of the Operational Research Society*, 60(4):481–495, 2009.

[304] F.S. Hillier and G.J. Lieberman. *Introduction to operations research*. 9th edition, McGraw-Hill, New York, NY, USA, 2009.

[305] J. Hillston. *A compositional approach to performance modelling*. Cambridge University Press, Cambridge, UK, 1996.

[306] J. Hillston and N. Thomas. A syntactic analysis of reversible PEPA processes. In *Proceedings 6th International Workshop on Process Algebra and Performance Modelling*, pages 37–50, 1998.

[307] J. Hillston and N. Thomas. Product form solution for a class of PEPA models. *Performance Evaluation*, 35(3-4):171–192, 1999.

[308] C.J. Ho and H.S. Lau. Minimizing total cost in scheduling outpatient appointments. *Management Science*, 38(12):1750–1764, 1992.

[309] C.J. Ho and H.S. Lau. Evaluating the impact of operating conditions on the performance of appointment scheduling rules in service systems. *European Journal of Operational Research*, 112(3):542–553, 1999.

[310] A. Hordijk and N.M. van Dijk. Networks of queues. *Modelling and Performance Evaluation Methodology*, 60:151–205, 1984.

[311] M.T.T. Hsiao and A.A. Lazar. An extension to Norton's equivalent. *Queueing Systems*, 5(4):401–411, 1989.

[312] V.N. Hsu, R. de Matta, and C.Y. Lee. Scheduling patients in an ambulatory surgical center. *Naval Research Logistics*, 50(3):218–238, 2003.

[313] R. Huang, S. Kim, and M.B.C. Menezes. Facility location for large-scale emergencies. *Annals of Operations Research*, 181(1):271–286, 2010.

[314] X.M. Huang. A planning model for requirement of emergency beds. *Mathematical Medicine and Biology*, 12(3-4):345, 1995.

[315] X.M. Huang. Decision making support in reshaping hospital medical services. *Health Care Management Science*, 1(2):165–173, 1998.

[316] M. Hughes, E.R. Carson, M. Makhlouf, C.J. Morgan, and R. Summers. Modelling a progressive care system using a coloured-timed Petri net. *Transactions of the Institute of Measurement & Control*, 22(3):271, 2000.

[317] W.L. Hughes and S.Y. Soliman. Short-term case mix management with linear programming. *Hospital & Health Services Administration*, 30(1):52–60, 1985.

[318] T. Huisman and R.J. Boucherie. Decomposition and aggregation in queueing networks. In R.J. Boucherie and N.M. van Dijk, editors, *Queueing networks: a fundamental approach*, pages 313–344. Springer, New York, NY, USA, 2010.

[319] P.J.H. Hulshof, R.J. Boucherie, J.T. van Essen, E.W. Hans, J.L. Hurink, N. Kortbeek, N. Litvak, P.T. Vanberkel, E. van der Veen, B. Veltman, I.M.H. Vliegen, and M.E. Zonderland. ORchestra: an online reference database of OR/MS literature in health care. *Health Care Management Science*, 14(4):383–384, 2011.

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

[321] P.J.H. Hulshof, P.T. Vanberkel, R.J. Boucherie, E.W. Hans, M. van Houdenhoven, and J.C.W. van Ommeren. Analytical models to determine room requirements in outpatient clinics. *OR Spectrum*, 34(2-SI):391–405, 2012.

[322] A. Ingolfsson, S. Budge, and E. Erkut. Optimal ambulance location with random delays and travel times. *Health Care Management Science*, 11(3):262–274, 2008.

[323] A. Ingolfsson, E. Erkut, and S. Budge. Simulation of single start station for Edmonton EMS. *Journal of the Operational Research Society*, 54(7):736–746, 2003.

[324] Institute for Operations Research and the Management Sciences (INFORMS). About Operations Research. *Retrieved October 13, 2012, from http://www.informs.org/About-INFORMS/About-Operations-Research*.

[325] Institute of Medicine. *Crossing the quality chasm: a new health system for the 21st century*. National Academies Press, Washington, DC, USA, 2001.

[326] V. Irvine, S.I. McClean, and P.H. Millard. Stochastic models for geriatric in-patient behaviour. *Mathematical Medicine and Biology*, 11(3):207–216, 1994.

[327] M.W. Isken, T.J. Ward, and S.J. Littig. An open source software project for obstetrical procedure scheduling and occupancy analysis. *Health Care Management Science*, 14(1):56–73, 2011.

[328] E.P. Jack and T.L. Powers. A review and synthesis of demand management, capacity management and performance in health-care services. *International Journal of Management Reviews*, 11(2):149–174, 2009.

[329] E.P. Jack and T.L. Powers. Volume flexible strategies in health services: a research framework. *Production and Operations Management*, 13(3):230–244, 2009.

[330] J.R. Jackson. Networks of waiting lines. *Operations Research*, 5(4):518–521, 1957.

[331] V. Jain, S. Wadhwa, and S.G. Deshmukh. Modelling and analysis of supply chain dynamics: a High Intelligent Time (HIT) Petri net based approach. *International Journal of Industrial and Systems Engineering*, 1(1):59–86, 2006.

[332] B. Jaumard, F. Semet, and T. Vovor. A generalized linear programming model for nurse scheduling. *European Journal of Operational Research*, 107(1):1–18, 1998.

[333] A. Jebali, H. Alouane, B. Atidel, and P. Ladet. Operating rooms scheduling. *International Journal of Production Economics*, 99(1-2):52–62, 2006.

[334] K. Jensen and L.M. Kristensen. *Coloured Petri nets: modelling and validation of concurrent systems*. Springer, Berlin, Germany, 2009.

[335] P.A. Jensen and J.F. Bard. *Operations Research models and methods*. John Wiley & Sons, London, UK, 2003.

[336] H. Jia, F. Ordóñez, and M. Dessouky. A modeling framework for facility location of medical services for large-scale emergencies. *IIE Transactions*, 39(1):41–55, 2007.

[337] P.E. Joustra, J. de Wit, V.M.D. Struben, B.J.H. Overbeek, P. Fockens, and S.G. Elkhuizen. Reducing access times for an endoscopy department by an iterative combination of computer simulation and Linear Programming. *Health Care Management Science*, 13(1):17–26, 2010.

[338] P.E. Joustra, J. de Wit, N.M. van Dijk, and P.J.M. Bakker. How to juggle priorities? an interactive tool to provide quantitative support for strategic patient-mix decisions: an ophthalmology case. *Health Care Management Science*, 14(4):348–360, 2011.

[339] J.B. Jun, S.H. Jacobson, and J.R. Swisher. Application of discrete-event simulation in health care clinics: a survey. *Journal of the Operational Research Society*, 50(2):109–123, 1999.

[340] G.C. Kaandorp and G. Koole. Optimal outpatient appointment scheduling. *Health Care Management Science*, 10(3):217–229, 2007.

[341] R.L. Kane, T.A. Shamliyan, C. Mueller, S. Duval, and T.J. Wilt. The association of registered nurse staffing levels and patient outcomes: systematic review and meta-analysis. *Medical Care*, 45(12):1195–1204, 2007.

[342] E.P.C. Kao and G.G. Tung. Bed allocation in a public health care delivery system. *Management Science*, 27(5):507–520, 1981.

[343] A.S. Kapadia and Y.K.C.M. Fasihullah. Finite capacity priority queues with potential health applications. *Computers & Operations Research*, 12(4):411–420, 1985.

[344] K. Katsaliaki and N. Mustafee. Applications of simulation within the healthcare context. *Journal of the Operational Research Society*, 62(8):1431–1451, 2010.

[345] J.H. Katz. Simulation of outpatient appointment systems. *Communications of the ACM*, 12(4):215–222, 1969.

[346] D.L. Kellogg and S. Walczak. Nurse scheduling: from academia to implementation or not? *Interfaces*, 37(4):355, 2007.

[347] F.P. Kelly. *Reversibility and stochastic networks*. New York, NY, USA. John Wiley & Sons, 1979.

[348] S.C. Kim and I. Horowitz. Scheduling hospital services: the efficacy of elective-surgery quotas. *Omega*, 30(5):335–346, 2002.

[349] S.C. Kim, I. Horowitz, K.K. Young, and T.A. Buckley. Analysis of capacity management of the intensive care unit in a hospital. *European Journal of Operational Research*, 115(1):36–46, 1999.

[350] S.C. Kim, I. Horowitz, K.K. Young, and T.A. Buckley. Flexible bed allocation and performance in the intensive care unit. *Journal of Operations Management*, 18(4):427–443, 2000.

[351] K.J. Klassen and T.R. Rohleder. Scheduling outpatient appointments in a dynamic environment. *Journal of Operations Management*, 14(2):83–101, 1996.

[352] K.J. Klassen and T.R. Rohleder. Outpatient appointment scheduling with urgent clients in a dynamic, multi-period environment. *International Journal of Service Industry Management*, 15(2):167–186, 2004.

[353] L. Kleinrock. *Queueing systems, volume 1: theory*. John Wiley & Sons, London, UK, 1975.

[354] N. Koizumi, E. Kuno, and T.E. Smith. Modeling patient flows using a queuing network with blocking. *Health Care Management Science*, 8(1):49–60, 2005.

[355] A. Kokangul. A combination of deterministic and stochastic approaches to optimize bed capacity in a hospital unit. *Computer Methods and Programs in Biomedicine*, 90(1):56–65, 2008.

[356] P. Kolesar. A Markovian model for hospital admission scheduling. *Management Science*, 16(6):384–396, 1970.

[357] R. Kolisch and S. Sickinger. Providing radiology health care services to stochastic demand of different customer classes. *OR Spectrum*, 30(2):375–395, 2008.

[358] R. Konrad, M. Lawley, and M. Criswell. Incorporating diagnosis based patient flow paths from health information systems messages into hospital decision models. In *Proceedings of the 3th INFORMS Workshop on Data Mining and Health Informatics*, pages 1–6, 2008.

[359] R. Kopach, P.C. DeLaurentis, M. Lawley, K. Muthuraman, L. Ozsen, R. Rardin, H. Wan, P. Intrevado, X. Qu, and D. Willis. Effects of clinical characteristics on successful open access scheduling. *Health Care Management Science*, 10(2):111–124, 2007.

[360] R. Kopach-Konrad, M. Lawley, M. Criswell, I. Hasan, S. Chakraborty, J. Pekny, and B.N. Doebbeling. Applying systems engineering principles in improving health care delivery. *Journal of General Internal Medicine*, 22(3):431–437, 2007.

[361] P.S. Kritzinger, S. Van Wyk, and A.E. Krzesinski. A generalisation of Norton's theorem for multiclass queueing networks. *Performance Evaluation*, 2(2):98–107, 1982.

[362] T. Kroll and M. Neri. Experiences with care co-ordination among people with cerebral palsy, multiple sclerosis, or spinal cord injury. *Disability & Rehabilitation*, 25(19):1106–1114, 2003.

[363] J. Kros, S. Dellana, and D. West. Overbooking increases patient access at East Carolina University's student health services clinic. *Interfaces*, 39(3):271–287, 2009.

[364] F. Krückeberg and M. Jaxy. Mathematical methods for calculating invariants in Petri nets. In G. Rozenberg, editor, *Advances in Petri nets*, volume 266 of *Lecture Notes in Computer Science*, pages 104–131. 1987.

[365] P.J. Kuzdrall, N.K. Kwak, and H.H. Schmitz. Simulating space requirements and scheduling policies in a hospital surgical suite. *Simulation*, 36(5):163–172, 1981.

[366] N.K. Kwak, P.J. Kuzdrall, and H.H. Schmitz. Simulating the use of space in a hospital surgical suite. *Simulation*, 25(5):147–151, 1975.

[367] N.K. Kwak, P.J. Kuzdrall, and H.H. Schmitz. The GPSS simulation of scheduling policies for surgical patients. *Management Science*, 22(9):982–989, 1976.

[368] L.R. LaGanga and S.R. Lawrence. Clinic overbooking to improve patient access and increase provider productivity. *Decision Sciences*, 38(2):251–276, 2007.

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

[370] N. Lahrichi, S.D. Lapierre, A. Hertz, A. Talib, and L. Bouvier. Analysis of a territorial approach to the delivery of nursing home care services based on historical data. *Journal of Medical Systems*, 30(4):283–291, 2006.

[371] 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.

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

[373] M. Lamiri, X. Xie, and S. Zhang. Column generation approach to operating theater planning with elective and emergency patients. *IIE Transactions*, 40(9):838–852, 2008.

[374] T.P. Landau, T.R. Thiagarajan, and R.S. Ledley. Cost containment in the concentrated care center: a study of nursing, bed and patient assignment policies. *Computers in Biology and Medicine*, 13(3):205–238, 1983.

[375] D.C. Lane and E. Husemann. System dynamics mapping of acute patient flows. *Journal of the Operational Research Society*, 59(2):213–224, 2007.

[376] D.C. Lane, C. Monefeldt, and J.V. Rosenhead. Looking in the wrong place for healthcare improvements: a system dynamics study of an accident and emergency department. *Journal of the Operational Research Society*, 51(5):518–531, 2000.

[377] T.A. Lang, M. Hodge, V. Olson, P.S. Romano, and R.L. Kravitz. Nurse-patient ratios: a systematic review on the effects of nurse staffing on patient, nurse employee, and hospital outcomes. *Journal of Nursing Administration*, 34(7-8):326–337, 2004.

[378] J.R. Langabeer. *Health care operations management: a quantitative approach to business and logistics*. Jones & Bartlett Publishers, Sudbury, MA, USA, 2007.

[379] E. Lanzarone, A. Matta, and G. Scaccabarozzi. A patient stochastic model to support human resource planning in home care. *Production Planning and Control*, 21(1):3–25, 2010.

[380] R.C. Larson. Approximating the performance of urban emergency service systems. *Operations Research*, 23(5):845–868, 1975.

[381] V. Lattimer, S. Brailsford, J. Turnbull, P. Tarnaras, H. Smith, S. George, K. Gerard, and S. Maslin-Prothero. Reviewing emergency care systems I: insights from system dynamics modelling. *Emergency Medicine Journal*, 21(6):685–691, 2004.

[382] M.S. Lavieri and M.L. Puterman. Optimizing nursing human resource planning in British Columbia. *Health Care Management Science*, 12(2):119–128, 2009.

[383] A.M. Law. *Simulation modeling and analysis*. McGraw-Hill, Boston, MA, USA, 4th edition, 2006.

[384] A.A. Lazar and T.G. Robertazzi. Markovian Petri net protocols with product form solution. In *Proceedings of the International Conference on Computer Communications*, pages 1054–1062, 1987.

[385] A.A. Lazar and T.G. Robertazzi. Markovian Petri net protocols with product form solution. *Performance Evaluation*, 12(1):67–77, 1991.

[386] D.K.K. Lee and S.A. Zenios. Optimal capacity overbooking for the regular treatment of chronic conditions. *Operations Research*, 57(4):852–865, 2009.

[387] D.Y. Lee and F. DiCesare. Scheduling flexible manufacturing systems using Petri nets and heuristic search. *IEEE Transactions on Robotics and Automation*, 10(2):123–132, 1994.

[388] S. Lee. The role of preparedness in ambulance dispatching. *Journal of the Operational Research Society*, 62(10):1888–1897, 2010.

[389] H.S. Leff, M. Dada, and S.C. Graves. An LP planning model for a mental health community support system. *Management Science*, 32(2):139–155, 1986.

[390] B. Lehaney, S.A. Clarke, and R.J. Paul. A case of an intervention in an outpatients department. *Journal of the Operational Research Society*, 50(9):877–891, 1999.

[391] C.R.M. Leite, D.L. Martin, G.R.M.A. Sizilio, K.E.A. Dos Santos, B.G. De Araujo, R.A.M. Valentim, A.D.D. Neto, J.D. De Melo, and A.M.G. Guerreiro. Modeling of medical care with stochastic petri nets. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1336–1339, 2010.

[392] L.L.X. Li and B.E. King. A healthcare staff decision model considering the effects of staff cross-training. *Health Care Management Science*, 2(1):53–61, 1999.

[393] L.X. Li and W.C. Benton. Performance measurement criteria in health care organizations: Review and future research directions. *European Journal of Operational Research*, 93(3):449–468, 1996.

[394] X. Li, P. Beullens, D. Jones, and M. Tamiz. An integrated queuing and multi-objective bed allocation model with application to a hospital in china. *Journal of the Operational Research Society*, 60(3):330–338, 2009.

[395] X. Li, Z. Zhao, X. Zhu, and T. Wyatt. Covering models and optimization techniques for emergency response facility location and planning: a review. *Mathematical Methods of Operations Research*, 74(3):1–30, 2011.

[396] C.J. Liao, C.D. Pegden, and M. Rosenshine. Planning timely arrivals to a stochastic production or service system. *IIE Transactions*, 25(5):63–73, 1993.

[397] C.S. Lim, R. Mamat, and T. Bräunl. Impact of ambulance dispatch policies on performance of emergency medical services. *IEEE Transactions on Intelligent Transportation Systems*, 12(2):624–632, 2011.

[398] S.J. Littig and M.W. Isken. Short term hospital occupancy prediction. *Health Care Management Science*, 10(1):47–66, 2007.

[399] E. Litvak, P.I. Buerhaus, F. Davidoff, M.C. Long, M.L. McManus, and D.M. Berwick. Managing unnecessary variability in patient demand to reduce nursing stress and improve patient safety. *Joint Commission Journal on Quality and Patient Safety*, 31(6):330–338, 2005.

[400] L. Liu and X. Liu. Block appointment systems for outpatient clinics with multiple doctors. *Journal of the Operational Research Society*, 49(12):1254–1259, 1998.

[401] L. Liu and X. Liu. Dynamic and static job allocation for multi-server systems. *IIE transactions*, 30(9):845–854, 1998.

[402] N. Liu, S. Ziya, and V.G. Kulkarni. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing & Service Operations Management*, 12(2):347–364, 2010.

[403] W.S. Lovejoy and Y. Li. Hospital operating room capacity expansion. *Management Science*, 48(11):1369–1387, 2002.

[404] M. Mackay. Practical experience with bed occupancy management and planning systems: an Australian view. *Health Care Management Science*, 4(1):47–56, 2001.

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

[406] J. Mairesse and H.T. Nguyen. Deficiency zero Petri nets and product form. *Fundamenta Informaticae*, 105(3):237–261, 2010.

[407] V. Maizes, D. Rakel, and C. Niemiec. Integrative medicine and patient-centered care. *Explore: The Journal of Science and Healing*, 5(5):277–289, 2009.

[408] R.S. Mans, H. Schonenberg, G. Leonardi, S. Panzarasa, A. Cavallini, S. Quaglini, and W.M.P. van der Aalst. Process mining techniques: an application to stroke care. *Studies in Health Technology and Informatics*, 136:573–578, 2008.

[409] R.S. Mans, W.M.P. Van Der Aalst, N.C. Russell, P.J.M. Bakker, A.J. Moleman, K.B. Lassen, and J.B. Jørgensen. From requirements via colored workflow nets to an implementation in several workflow systems. In K. Jensen, J. Billington, and M. Koutny, editors, *Transactions on Petri nets and other models of concurrency III*, number 5800 in Lecture Notes in Computer Science, pages 25–49. 2009.

[410] A. Manzano-Santaella. From bed-blocking to delayed discharges: precursors and interpretations of a contested concept. *Health Services Management Research*, 23(3):121–127, 2010.

[411] E. Marcon and F. Dexter. Impact of surgical sequencing on post anesthesia care unit staffing. *Health Care Management Science*, 9(1):87–98, 2006.

[412] 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.

[413] V. Marianov and C. ReVelle. The queueing maximal availability location problem: a model for the siting of emergency vehicles. *European Journal of Operational Research*, 93(1):110–120, 1996.

[414] A. Marin. *On the relations among product-form stochastic models*. PhD thesis, Universita Ca' Foscari di Venezia, Italy, 2009.

[415] R.A. Marjamaa, P.M. Torkki, E.J. Hirvensalo, and O.A. Kirvelä. What is the best workflow for an operating room? A simulation study of five scenarios. *Health Care Management Science*, 12(2):142–146, 2009.

[416] I. Marques, M.E. Captivo, and M. Vaz Pato. An integer programming approach to elective surgery scheduling. *OR Spectrum*, 34(2):407–427, 2011.

[417] A.M. Marsan, A. Bobbio, and S. Donatelli. Petri nets in performance analysis: an introduction. In W. Reisig and G. Rozenberg, editors, *Lectures on Petri nets I: basic models*, volume 1491 of *Lecture Notes in Computer Science*, pages 211–256. 1998.

[418] M.A. Marsan, G. Balbo, A. Bobbio, G. Chiola, G. Conte, and A. Cumani. The effect of execution policies on the semantics and analysis of stochastic Petri nets. *IEEE Transactions on Software Engineering*, 15(7):846, 1989.

[419] M.A. Marsan, G. Balbo, G. Conte, S. Donatelli, and G. Franceschinis. *Modelling with generalized stochastic Petri nets*. John Wiley & Sons, New York, NY, USA, 1994.

[420] A.H. Marshall and S.I. McClean. Using coxian phase-type distributions to identify patient characteristics for duration of stay in hospital. *Health Care Management Science*, 7(4):285–289, 2004.

[421] A.H. Marshall, S.I. McClean, and P.H. Millard. Addressing bed costs for the elderly: a new methodology for modelling patient outcomes and length of stay. *Health Care Management Science*, 7(1):27–33, 2004.

[422] A.H. Marshall, S.I. McClean, C.M. Shapcott, and P.H. Millard. Modelling patient duration of stay to facilitate resource management of geriatric hospitals. *Health Care Management Science*, 5(4):313–319, 2002.

[423] A.H. Marshall, B. Shaw, and S.I. McClean. Estimating the costs for a group of geriatric patients using the coxian phase-type distribution. *Statistics in Medicine*, 26(13):2716–2729, 2007.

[424] B.J. Masterson, T.G. Mihara, G. Miller, S.C. Randolph, M.E. Forkner, and A.L. Crouter. Using models and data to support optimization of the military health system: a case study in an intensive care unit. *Health Care Management Science*, 7(3):217–224, 2004.

[425] M.E. Matta and S.S. Patterson. Evaluating multiple performance measures across several dimensions at a multi-facility outpatient center. *Health Care Management Science*, 10(2):173–194, 2007.

[426] R.S. Maull, P.A. Smart, A. Harris, and A.A.F. Karasneh. An evaluation of 'fast track' in A&E: a discrete event simulation approach. *The Service Industries Journal*, 29(7):923–941, 2009.

[427] M.S. Maxwell, M. Restrepo, S.G. Henderson, and H. Topaloglu. Approximate dynamic programming for ambulance redeployment. *INFORMS Journal on Computing*, 22(2):266–281, 2010.

[428] J.H. May, W.E. Spangler, D.P. Strum, and L.G. Vargas. The surgical scheduling problem: Current research and future opportunities. *Production and Operations Management*, 20(3):392–405, 2011.

[429] L. Mayhew and D. Smith. Using queuing theory to analyse the government's 4-h completion time target in accident and emergency departments. *Health Care Management Science*, 11(1):11–21, 2008.

[430] J.O. McClain. A model for regional obstetric bed planning. *Health Services Research*, 13(4):378–394, 1978.

[431] S.I. McClean, M. Barton, L. Garg, and K. Fullerton. A modeling framework that combines markov models and discrete-event simulation for stroke patient care. *ACM Transactions on Modeling and Computer Simulation*, 21(4):25, 2011.

[432] S.I. McClean, B. McAlea, and P.H. Millard. Using a Markov reward model to estimate spend-down costs for a geriatric department. *Journal of the Operational Research Society*, 49(10):1021–1025, 1998.

[433] S.I. McClean and P.H. Millard. Patterns of length of stay after admission in geriatric medicine: an event history approach. *The Statistician*, 42(3):263–274, 1993.

[434] S.I. McClean and P.H. Millard. A three compartment model of the patient flows in a geriatric department: a decision support approach. *Health Care Management Science*, 1(2):159–163, 1998.

[435] S.I. McClean and P.H. Millard. Where to treat the older patient? Can Markov models help us better understand the relationship between hospital and community care? *Journal of the Operational Research Society*, 58(2):255–261, 2007.

[436] C. McLaughlin. Why variation reduction is not everything: a new paradigm for service operations. *International Journal of Service Industry Management*, 7(3):17–30, 1996.

[437] D.B. McLaughlin and J.M. Hays. *Healthcare operations management*. Health Administration Press, Chicago, IL, USA, 2008.

[438] M.L. McManus, M.C. Long, A. Cooper, J. Mandell, D.M. Berwick, M. Pagano, and E. Litvak. Variability in surgical caseload and access to intensive care services. *Anesthesiology*, 98(6):1491–1496, 2003.

[439] Medline Plus. Neuromuscular disorders. *Retrieved October 13, 2012, from http://www.nlm.nih.gov/medlineplus/neuromusculardisorders.html*.

[440] G. Memmi and G. Roucairol. Linear algebra in net theory. In *Proceedings of the Advanced Course on General Net Theory of Processes and Systems: Net Theory and Applications*, pages 213–223, 1980.

[441] B. Mielczarek and J. Uziałko-Mydlikowska. Application of computer simulation modeling in the health care sector: a survey. *Simulation*, 88(2):197–216, 2012.

[442] P.H. Millard, G. Christodoulou, C. Jagger, G.W. Harrison, and S.I. McClean. Modelling hospital and social care bed occupancy and use by elderly people in an English health district. *Health Care Management Science*, 4(1):57–62, 2001.

[443] D. Min and Y. Yih. An elective surgery scheduling problem considering patient priority. *Computers & Operations Research*, 37(6):1091–1099, 2010.

[444] D. Min and Y. Yih. Scheduling elective surgery under uncertainty and downstream capacity constraints. *European Journal of Operational Research*, 206(3):642–652, 2010.

[445] Ministerie van Volksgezondheid, Welzijn en Sport [Ministry of Health, Welfare and Sport]. De zorg: hoeveel extra is het ons waard? [Healthcare: how much more is it worth us?]. *Retrieved October 13, 2012, from http://www.rijksoverheid.nl/documenten-en-publicaties/rapporten/2012/06/12/rapport-de-zorg-hoeveel-extra-is-het-ons-waard.html*, 2012.

[446] M. Miyazawa. Palm calculus, reallocatable gsmp and insensitivity structure. In R.J. Boucherie and N.M. Van Dijk, editors, *Queueing networks: a fundamental approach*, pages 141–216. Springer, New York, NY, USA, 2010.

[447] C. Mullinax and M. Lawley. Assigning patients to nurses in neonatal intensive care. *Journal of the Operational Research Society*, 53(1):25–35, 2002.

[448] T. Murata. Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, 77(4):541–580, 1989.

[449] M. Murray and D.M. Berwick. Advanced access: reducing waiting and delays in primary care. *Journal of the American Medical Association*, 289(8):1035–1040, 2003.

[450] N. Mustafee, K. Katsaliaki, and S.J.E. Taylor. Profiling literature in healthcare simulation. *Simulation*, 86(8-9):543, 2010.

[451] K. Muthuraman and M. Lawley. A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Transactions*, 40(9):820–837, 2008.

[452] National Biological Information Infrastructure (NBII). Homepage. *Retrieved September 22, 2011, from http://www.nbii.gov*.

[453] J. Needleman, P. Buerhaus, S. Mattke, M. Stewart, and K. Zelevinsky. Nurse-staffing levels and the quality of care in hospitals. *New England Journal of Medicine*, 346(22):1715–1722, 2002.

[454] M.F. Neuts. A general class of bulk queues with poisson input. *The Annals of Mathematical Statistics*, 38(3):759–770, 1967.

[455] M.F. Neuts. Queues solvable without Rouche's theorem. *Operations Research*, 27(4):767–781, 1979.

[456] H.T. Nguyen. *Reseaux de Petri stochastiques a forme produit [Stochastic Petri nets with product form]*. PhD thesis, Université Paris 7, France, 2012.

[457] J.M. Nguyen, P. Six, D. Antonioli, P. Glemain, G. Potel, P. Lombrail, and P. Le Beux. A simple method to optimize hospital beds capacity. *International Journal of Medical Informatics*, 74(1):39–49, 2005.

[458] J.M. Nguyen, P. Six, T.J. Chaussalet, D. Antonioli, P. Lombrail, and P. Le Beux. An objective method for bed capacity planning in a hospital department – a comparison with target ratio methods. *Methods of Information in Medicine*, 46(4):399–405, 2007.

[459] J.M. Nguyen, P. Six, R. Parisot, D. Antonioli, F. Nicolas, and P. Lombrail. A universal method for determining intensive care unit bed requirements. *Intensive Care Medicine*, 29(5):849–852, 2003.

[460] J.P. Oddoye, D.F. Jones, M. Tamiz, and P. Schmidt. Combining simulation and goal programming for healthcare planning in a medical assessment unit. *European Journal of Operational Research*, 193(1):250–261, 2009.

[461] J.P. Oddoye, M.A. Yaghoobi, M. Tamiz, D.F. Jones, and P. Schmidt. A multi-objective model to determine efficient resource levels in a medical assessment unit. *Journal of the Operational Research Society*, 58(12):1563–1573, 2007.

[462] S.N. Ogulata, M. Koyuncu, and E. Karakas. Personnel and patient scheduling in the high demanded hospital services: a case study in the physiotherapy service. *Journal of Medical Systems*, 32(3):221–228, 2008.

[463] H.C. Oh and W.L. Chow. Scientific evaluation of polyclinic operating strategies with discrete-event simulation. *International Journal of Simulation Modelling*, 10(4):165–176, 2011.

[464] M. Olivares, C. Terwiesch, and L. Cassorla. Structural estimation of the newsvendor model: an application to reserving operating room time. *Management Science*, 54(1):41–55, 2008.

[465] Organisation of Economic Co-operation and Development (OECD). OECD Glossary. *Retrieved October 13, 2012, from http://http://stats.oecd.org/glossary/*.

[466] Organisation of Economic Co-operation and Development (OECD). *Health at a glance 2011: OECD Indicators*. OECD Publishing, *Retreived October 13, 2012, from http://dx.doi.org/10.1787/health_glance-2011-en*, 2011.

[467] Y.A. Ozcan. *Health care benchmarking and performance evaluation: an assessment using Data Envelopment Analysis (DEA)*, volume 120 of *International Series in Operations Research & Management Science*. Springer, New York, NY, USA, 2007.

[468] Y.A. Ozcan. *Quantitative methods in health care management: techniques and applications*. Jossey Bass, San Francisco, CA, USA, 2nd edition, 2009.

[469] C.H. Papadimitriou and K. Steiglitz. *Combinatorial optimization: algorithms and complexity*. Dover Publications, Mineola, NY, USA, 1998.

[470] F. Pasin, M.H. Jobin, and J.F. Cordeau. An application of simulation to analyse resource sharing among health-care organisations. *International Journal of Operations and Production Management*, 22(4):381–393, 2002.

[471] J. Patrick. Access to long-term care: the true cause of hospital congestion? *Production and Operations Management*, 20(3):347–358, 2011.

[472] J. Patrick. A Markov decision model for determining optimal outpatient scheduling. *Health Care Management Science*, 15(2):91–102, 2012.

[473] J. Patrick and M.L. Puterman. Improving resource utilization for diagnostic services through flexible inpatient scheduling: a method for improving resource utilization. *Journal of the Operational Research Society*, 58(2):235–245, 2007.

[474] J. Patrick, M.L. Puterman, and M. Queyranne. Dynamic multi-priority patient scheduling for a diagnostic resource. *Operations Research*, 56(6):1507–1525, 2008.

[475] S.A. Paul, M.C. Reddy, and C.J. DeFlitch. A systematic review of simulation studies investigating emergency department overcrowding. *Simulation*, 86(8-9):559–571, 2010.

[476] C.D. Pegden and M. Rosenshine. Scheduling arrivals to queues. *Computers & Operations Research*, 17(4):343–348, 1990.

[477] C. Pelletier, T.J. Chaussalet, and H. Xie. A framework for predicting gross institutional long-term care cost arising from known commitments at local authority level. *Journal of the Operational Research Society*, 56(2):144–152, 2004.

[478] E. Pérez, L. Ntaimo, C. Bailey, and P. McCormack. Modeling and simulation of nuclear medicine patient service management in DEVS. *Simulation*, 86(8-9):481–501, 2010.

[479] M.J. Persson and J.A. Persson. Analysing management policies for operating room planning using simulation. *Health Care Management Science*, 13(2):182–191, 2010.

[480] J.L. Peterson. *Petri net theory and the modeling of systems*. Prentice Hall, Upper Saddle River, NJ, USA, 1981.

[481] 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.

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

[483] W.P. Pierskalla and D.J. Brailer. Applications of operations research in health care delivery. In S.M. Pollock, M.H. Rothkopf, and A. Barnett, editors, *Operations research and the public sector*, volume 6 of *Handbooks in OR & MS*, pages 469–505. North-Holland, Amsterdam, the Netherlands, 1994.

[484] 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.

[485] M.E. Porter. *Competitive advantage: creating and sustaining superior performance*. Free Press, New York, NY, USA, 1985.

[486] J.C. Prentice and S.D. Pizer. Delayed access to health care and mortality. *Health Services Management Research*, 42(2):644–662, 2007.

[487] C. Price, B. Golden, M. Harrington, R. Konewko, E. Wasil, and W. Herring. Reducing boarding in a post-anesthesia care unit. *Production and Operations Management*, 20(3):431–441, 2011.

[488] Z.H. Przasnyski. Operating room scheduling. a literature review. *Association of Perioperative Registered Nurses Journal*, 44(1):67–82, 1986.

[489] P. Punnakitikashem, J.M. Rosenberger, and D. Buckley Behan. Stochastic programming for nurse assignment. *Computational Optimization and Applications*, 40(3):321–349, 2008.

[490] X. Qu, R.L. Rardin, J.A.S. Williams, and D.R. Willis. Matching daily healthcare provider capacity to demand in advanced access scheduling systems. *European Journal of Operational Research*, 183(2):812–826, 2007.

[491] X. Qu and J. Shi. Effect of two-level provider capacities on the performance of open access clinics. *Health Care Management Science*, 12(1):99–114, 2009.

[492] X. Qu and J. Shi. Modeling the effect of patient choice on the performance of open access scheduling. *International Journal of Production Economics*, 129(2):314–327, 2011.

[493] S. Quaglini, E. Caffi, A. Cavallini, G. Micieli, and M. Stefanelli. Simulation of a stroke unit careflow. *Studies in Health Technology and Informatics*, 84(2):1190–1194, 2001.

[494] Raad voor Volksgezondheid & Zorg (RVZ) [Council for Public Health and Health Care]. Medisch-specialistische zorg in 2020 [Specialized medical care in 2020]. *Retrieved October 13, 2012, from http://http://www.rvz.net/*, 2011.

[495] A. Rais and A. Viana. Operations research in healthcare: a survey. *International Transactions in Operational Research*, 18(1):1–31, 2011.

[496] M. Ramakrishnan, D. Sier, and P.G. Taylor. A two-time-scale model for hospital patient flow. *IMA Journal of Management Mathematics*, 16(3):197, 2005.

[497] T.A. Reilly, V.P. Marathe, and B.E. Fries. A delay-scheduling model for patients using a walk-in clinic. *Journal of Medical Systems*, 2(4):303–313, 1978.

[498] J.F. Repede and J.J. Bernardo. Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *European Journal of Operational Research*, 75(3):567–581, 1994.

[499] C. ReVelle. Review, extension and prediction in emergency service siting models. *European Journal of Operational Research*, 40(1):58–69, 1989.

[500] J.C. Ridge, S.K. Jones, M.S. Nielsen, and A.K. Shahani. Capacity planning for intensive care units. *European Journal of Operational Research*, 105(2):346–355, 1998.

[501] A. Riise and E.K. Burke. Local search for the surgery admission planning problem. *Journal of Heuristics*, 17(4):389–414, 2011.

[502] E.J. Rising, R. Baron, and B. Averill. A systems analysis of a university-health-service outpatient clinic. *Operations Research*, 21(5):1030–1047, 1973.

[503] L.W. Robinson and R.R. Chen. Scheduling doctors' appointments: optimal and empirically-based heuristic policies. *IIE Transactions*, 35(3):295–307, 2003.

[504] L.W. Robinson and R.R. Chen. A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing Service Operations Management*, 12(2):330–346, 2010.

[505] T.R. Rohleder, D.P. Bischak, and L.B. Baskin. Modeling patient service centers with simulation and system dynamics. *Health Care Management Science*, 10(1):1–12, 2007.

[506] T.R. Rohleder, P. Lewkonia, D.P. Bischak, P. Duffy, and R. Hendijani. Using simulation modeling to improve patient flow at an outpatient orthopedic clinic. *Health Care Management Science*, 14(2):135–145, 2011.

[507] B. Roland, C. Di Martinelly, F. Riane, and Y. Pochet. Scheduling an operating theatre under human resource constraints. *Computers & Industrial Engineering*, 58(2):212–220, 2010.

[508] E. Rönnberg and T. Larsson. Automating the self-scheduling process of nurses in Swedish healthcare: a pilot study. *Health Care Management Science*, 13(1):35–53, 2010.

[509] S.M. Ross. *Stochastic processes*. John Wiley & Sons, New York, NY, USA, 2nd edition, 1996.

[510] S.M. Ross. *Introduction to probability models*. Elsevier Academic Press, Amsterdam, the Netherlands, 9th edition, 2007.

[511] R.J. Ruth. A mixed integer programming model for regional planning of a hospital inpatient service. *Management Science*, 7(5):521–533, 1981.

[512] K. Salimifard and M. Wright. Petri net-based modelling of workflow systems: an overview. *European Journal of Operational Research*, 134(3):664–676, 2001.

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

[514] E.S. Savas. Simulation and cost-effectiveness analysis of New York's emergency ambulance service. *Management Science*, 15(12):608–627, 1969.

[515] W. Schäfer, M. Kroneman, W. Boerma, M. van den Berg, G. Westert, W. Devillé, and E. van Ginneken. The Netherlands: health system review. *Health Systems in Transition*, 12(1):1–229, 2010.

[516] J. Scheer, T. Kroll, M.T. Neri, and P. Beatty. Access barriers for persons with disabilities. *Journal of Disability Policy Studies*, 13(4):221–230, 2003.

[517] K. Schimmelpfeng, S. Helber, and S. Kasper. Decision support for rehabilitation hospital scheduling. *OR Spectrum*, 32(2):1–29, 2010.

[518] V. Schmid. Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. *European Journal of Operational Research*, 219(3):611–621, 2012.

[519] H.H. Schmitz and N.K. Kwak. Monte Carlo simulation of operating-room and recovery-room usage. *Operations Research*, 20(6):1171–1180, 1972.

[520] H.H. Schmitz, N.K. Kwak, and P.J. Kuzdrall. Determination of surgical suite capacity and an evaluation of patient scheduling policies. *RAIRO - Operations Research - Recherche Opérationnelle*, 12(1):3–14, 1978.

[521] A. Schrijver. *Combinatorial optimization: polyhedra and efficiency*. Springer, New York, NY, USA, 2003.

[522] SEO Economisch Onderzoek [SEO Economic Research]. Kosten en baten van revalidatie [Costs and benefits of rehabilitation]. *Retreived October 13, 2012, from http://www.revalidatienederland.nl/actueel/publicaties*, 2008.

[523] M. Sereno. Towards a product form solution for stochastic process algebras. *The Computer Journal*, 38(7):622, 1995.

[524] M. Sereno and G. Balbo. Computational algorithms for product form solution stochastic Petri nets. In *Proceedings of the 5th International Workshop on Petri Nets and Performance Models*, pages 98–107, 1993.

[525] M. Sereno and G. Balbo. Mean value analysis of stochastic Petri nets. *Performance Evaluation*, 29(1):35–62, 1997.

[526] D.G. Seymour. Health care modelling and clinical practice. theoretical exercise or practical tool? *Health Care Management Science*, 4(1):7–12, 2001.

[527] A.K. Shahani, S.A. Ridley, and M.S. Nielsen. Modelling patient flows as an aid to decision making for critical care capacities and organisation. *Anaesthesia*, 63(10):1074–1080, 2008.

[528] B. Shaw and A.H. Marshall. Modeling the health care costs of geriatric inpatients. *IEEE Transactions on Information Technology in Biomedicine*, 10(3):526–532, 2006.

[529] S. Shepperd, J. McClaran, C.O. Phillips, N.A. Lannin, L.M. Clemson, A. McCluskey, I.D. Cameron, and S.L. Barras. Discharge planning from hospital to home. *Cochrane Database of Systematic Reviews*, 1:CD000313, 2010.

[530] A. Shmueli, C.L. Sprung, and E.H. Kaplan. Optimizing admissions to an intensive care unit. *Health Care Management Science*, 6(3):131–136, 2003.

[531] W. Shonick and J.R. Jackson. An improved stochastic model for occupancy-related random variables in general-acute hospitals. *Operations Research*, 21(4):952–965, 1973.

[532] S. Sickinger and R. Kolisch. The performance of a generalized Bailey–Welch rule for outpatient appointment scheduling under inpatient and emergency demand. *Health Care Management Science*, 12(4):408–419, 2009.

[533] D. Sier, P. Tobin, and C. McGurk. Scheduling surgical procedures. *Journal of the Operational Research Society*, 48(9):884–891, 1997.

[534] S.P. Siferd and W.C. Benton. Workforce staffing and scheduling: hospital nursing specific models. *European Journal of Operational Research*, 60(3):233–246, 1992.

[535] M. Singer and P. Donoso. Assessing an ambulance service with queuing theory. *Computers & Operations Research*, 35(8):2549–2560, 2008.

[536] D. Sinreich and O. Jabali. Staggered work shifts: a way to downsize and restructure an emergency department workforce yet maintain current operational performance. *Health Care Management Science*, 10(3):293–308, 2007.

[537] D. Sinreich, O. Jabali, and N.P. Dellaert. Reducing emergency department waiting times by adjusting work shifts considering patient visits to multiple care providers. *IIE Transactions*, 44(3):163–180, 2012.

[538] C.E. Smith, S.V.M. Kleinbeck, K. Fernengel, and L.S. Mayer. Efficiency of families managing home health care. *Annals of Operations Research*, 73:157–175, 1997.

[539] K.R. Smith, A.M. Over Jr., M.F. Hansen, F.L. Golladay, and E.J. Davenport. Analytic framework and measurement strategy for investigating optimal staffing in medical practice. *Operations Research*, 24(5):815–841, 1976.

[540] V.L. Smith-Daniels, S.B. Schweikhart, and D.E. Smith-Daniels. Capacity management in health care services: review and future research directions. *Decision Sciences*, 19(4):889–919, 1988.

[541] B.G. Sobolev, V. Sanchez, and C. Vasilakis. Systematic review of the use of computer simulation modeling of patient flow in surgical care. *Journal of Medical Systems*, 35(1):1–16, 2011.

[542] A. Sonnenberg. How to overbook procedures in the endoscopy unit. *Gastrointestinal Endoscopy*, 69(3-Part-2):710–715, 2009.

[543] E.F. Stafford Jr. and S.C. Aggarwal. Managerial analysis and decision-making in outpatient health clinics. *Journal of the Operational Research Society*, 30(10):905–915, 1979.

[544] M.C. Su, S.C. Lai, P.C. Wang, Y.Z. Hsieh, and S.C. Lin. A SOMO-based approach to the operating room scheduling problem. *Expert Systems with Applications*, 38(12):15447–15454, 2011.

[545] S. Su and C.L. Shih. Managing a mixed-registration-type appointment system in outpatient clinics. *International Journal of Medical Informatics*, 70(1):31–40, 2003.

[546] D. Sundaramoorthi, V.C.P. Chen, J.M. Rosenberger, S.B. Kim, and D.F. Buckley-Behan. A data-integrated simulation model to evaluate nurse–patient assignments. *Health Care Management Science*, 12(3):252–268, 2009.

[547] J.R. Swisher and S.H. Jacobson. Evaluating the design of a family practice healthcare clinic using discrete-event simulation. *Health Care Management Science*, 5(2):75–88, 2002.

[548] J.R. Swisher, S.H. Jacobson, J.B. Jun, and O. Balci. Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation. *Computers and Operations Research*, 28(2):105–125, 2001.

[549] C. Swoveland, D. Uyeno, I. Vertinsky, and R. Vickson. Ambulance location: a probabilistic enumeration approach. *Management Science*, 20(4):686–698, 1973.

[550] H.A. Taha. *Operations research: an introduction*. Prentice Hall, Upper Saddle River, NJ, USA, 9th edition, 2010.

[551] H. Takagi. Queuing analysis of polling models. *ACM Computing Surveys*, 20(1):5–28, 1988.

[552] E. Tànfani and A. Testi. A pre-assignment heuristic algorithm for the Master Surgical Schedule Problem (MSSP). *Annals of Operations Research*, 178(1):105–119, 2010.

[553] H. Tarakci, Z. Ozdemir, and M. Sharafali. On the staffing policy and technology investment in a specialty hospital offering telemedicine. *Decision Support Systems*, 46(2):468–480, 2009.

[554] E. Tavares, J. Aleixo, P. Maciel, D. Oliveira, E. Heyde, R. Araujo, L. Maia, A. Duarte, and M. Novaes. Performance evaluation of medical imaging service. In *Proceedings of the 27th Annual Association for Computing Machinery Symposium on Applied Computing*, pages 1349–1354, 2012.

[555] G.J. Taylor, S. McClean, and P.H. Millard. Geriatric-patient flow-rate modelling. *Mathematical Medicine and Biology*, 13(4):297, 1996.

[556] G.J. Taylor, S.I. McClean, and P.H. Millard. Continuous-time markov models for geriatric patient behaviour. *Applied Stochastic Models and Data Analysis*, 13(3-4):315–323, 1997.

[557] G.J. Taylor, S.I. McClean, and P.H. Millard. Using a continuous-time Markov model with Poisson arrivals to describe the movements of geriatric patients. *Applied Stochastic Models and Data Analysis*, 14(2):165–174, 1998.

[558] G.J. Taylor, S.I. McClean, and P.H. Millard. Stochastic models of geriatric patient bed occupancy behaviour. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(1):39–48, 2000.

[559] I.D.S. Taylor and J.G.C. Templeton. Waiting time in a multi-server cutoff-priority queue, and its application to an urban ambulance service. *Operations Research*, 28(5):1168–1188, 1980.

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

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

[562] S.J. Thomas. Capacity and demand models for radiotherapy treatment machines. *Clinical Oncology*, 15(6):353–358, 2003.

[563] S. Thompson, M. Nunez, R. Garfinkel, and M.D. Dean. Efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges. *Operations Research*, 57(2):261–273, 2009.

[564] Thomson Reuters. Web of Science (WoS). *Retrieved October 13, 2012, from http://www.isiknowledge.com/*.

[565] H.C. Tijms. *A first course in stochastic models*. John Wiley & Sons, New York, NY, USA, 2003.

[566] C. Toregas, R. Swain, C. ReVelle, and L. Bergman. The location of emergency service facilities. *Operations Research*, 19(6):1363–1373, 1971.

[567] A. Trautsamwieser, M. Gronalt, and P. Hirsch. Securing home health care in times of natural disasters. *OR Spectrum*, 33(3):1–27, 2011.

[568] Treeknorm. *Retrieved October 13, 2012, from http://www.treeknorm.nl/*.

[569] V.M. Trivedi and D.M. Warner. A branch and bound algorithm for optimum allocation of float nurses. *Management Science*, 22(9):972–981, 1976.

[570] P.M. Troy and L. Rosenberg. Using simulation to determine the need for ICU beds for surgery patients. *Surgery*, 146(4):608–620, 2009.

[571] W.J.C. Tunnicliffe. A review of operational problems tackled by computer simulation in health care facilities. *Health and Social Service Journal*, 90(4702):73–80, 1980.

[572] A. Turkcan, B. Zeng, and M. Lawley. Chemotherapy operations planning and scheduling. *IIE Transactions on Healthcare Systems Engineering*, 2(1):31–49, 2012.

[573] D. Twigg, C. Duffield, A. Bremner, P. Rapley, and J. Finn. The impact of the nursing hours per patient day (NHPPD) staffing method on patient outcomes: a retrospective analysis of patient and staffing data. *International Journal of Nursing Studies*, 48(5):540–548, 2011.

[574] U.S. National Library of Medicine. Medical Subject Headings (MeSH). *Retrieved October 13, 2012, from http://www.nlm.nih.gov/mesh/*.

[575] U.S. National Library of Medicine. Pubmed. *Retrieved October 13, 2012, from http://www.pubmed.gov/*.

[576] M. Utley, S. Gallivan, K. Davis, P. Daniel, P. Reeves, and J. Worrall. Estimating bed requirements for an intermediate care facility. *European Journal of Operational Research*, 150(1):92–100, 2003.

[577] M. Utley, S. Gallivan, T. Treasure, and O. Valencia. Analytical methods for calculating the capacity required to operate an effective booked admissions policy for elective inpatient services. *Health Care Management Science*, 6(2):97–104, 2003.

[578] M. Utley, M. Jit, and S. Gallivan. Restructuring routine elective services to reduce overall capacity requirements within a local health economy. *Health Care Management Science*, 11(3):240–247, 2008.

[579] C. Valouxis and E. Housos. Hybrid optimization techniques for the workshift and rest assignment of nursing personnel. *Artificial Intelligence in Medicine*, 20(2):155–175, 2000.

[580] W.M.P. van der Aalst. The application of Petri nets to workflow management. *Journal of Circuits Systems and Computers*, 8(1):21–66, 1998.

[581] W.M.P. van der Aalst and B. van Dongen. Discovering workflow performance models from timed logs. In Y. Han, S. Tai, and D. Wikarski, editors, *Engineering and deployment of cooperative information systems*, volume 2480 of *Lecture Notes in Computer Science*, pages 107–110. 2002.

[582] N.M. van Dijk. *Queueing networks and product forms: a systems approach*. John Wiley & Sons, Chichester, UK, 1993.

[583] N.M. van Dijk and N. Kortbeek. Erlang loss bounds for OT–ICU systems. *Queueing Systems*, 63(1):253–280, 2009.

[584] B. van Dongen and W.M.P. van der Aalst. EMiT: a process mining tool. In J. Cortadella and W. Reisig, editors, *Applications and theory of Petri nets*, volume 3099 of *Lecture Notes in Computer Science*, pages 454–463. 2004.

[585] E. van Gameren and I. Woittiez. Transitions between care provisions demanded by Dutch elderly. *Health Care Management Science*, 8(4):299–313, 2005.

[586] K. van Hee, O. Oanea, R. Post, L. Somers, and J.M. van der Werf. Yasper: a tool for workflow modeling and analysis. In *Proceedings of the 6th International Conference on Application of Concurrency to System Design*, pages 279–282, 2006.

[587] M. van Houdenhoven, J.M. van Oostrum, E.W. Hans, G. Wullink, and G. Kazemier. Improving operating room efficiency by applying bin-packing and portfolio techniques to surgical case scheduling. *Anesthesia & Analgesia*, 105(3):707–714, 2007.

[588] M. van Houdenhoven, J.M. van Oostrum, G. Wullink, E.W. Hans, J.L. Hurink, J. Bakker, and G. Kazemier. Fewer intensive care unit refusals and a higher capacity utilization by using a cyclic surgical case schedule. *Journal of Critical Care*, 23(2):222–226, 2008.

[589] J.M. van Oostrum, M. van Houdenhoven, J.L. Hurink, E.W. Hans, G. Wullink, and G. Kazemier. A master surgical scheduling approach for cyclic scheduling in operating room departments. *OR Spectrum*, 30(2):355–374, 2008.

[590] P.T. Vanberkel and J.T. Blake. A comprehensive simulation for wait time reduction and capacity planning applied in general surgery. *Health Care Management Science*, 10(4):373–385, 2007.

[591] P.T. Vanberkel, R.J. Boucherie, E.W. Hans, J.L. Hurink, and N. Litvak. A survey of health care models that encompass multiple departments. *International Journal of Health Management and Information*, 1(1):37–69, 2010.

[592] P.T. Vanberkel, R.J. Boucherie, E.W. Hans, J.L. Hurink, W.A.M. van Lent, and W.H. van Harten. Accounting for inpatient wards when developing master surgical schedules. *Anesthesia & Analgesia*, 112(6):1472–1479, 2011.

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

[594] P.M.V. Vanden Bosch and D.C. Dietz. Minimizing expected waiting in a medical appointment system. *IIE Transactions*, 32(9):841–848, 2000.

[595] P.M.V. Vanden Bosch, D.C. Dietz, and J.R. Simeoni. Scheduling customer arrivals to a stochastic service system. *Naval Research Logistics*, 46(5):549–559, 1999.

[596] S. Vanderby and M.W. Carter. An evaluation of the applicability of system dynamics to patient flow modelling. *Journal of the Operational Research Society*, 61(11):1572–1581, 2009.

[597] K. Vanhaecht, K. de Witte, R. Depreitere, and W. Sermeus. Clinical pathway audit tools: a systematic review. *Journal of Nursing Management*, 14(7):529–537, 2006.

[598] C. Vasilakis and E. El-Darzi. A simulation study of the winter bed crisis. *Health Care Management Science*, 4(1):31–36, 2001.

[599] C. Vasilakis, B.G. Sobolev, L. Kuramoto, and A.R. Levy. A simulation study of scheduling clinic appointments in surgical care: individual surgeon versus pooled lists. *Journal of the Operational Research Society*, 58(2):202–211, 2007.

[600] G. Vassilacopoulos. A simulation model for bed allocation to hospital inpatient departments. *Simulation*, 45(5):233–241, 1985.

[601] Vereniging Spierziekten Nederland [Dutch Association of Neuromuscular Diseases]. Overzicht spierziekten [Overview neuromuscular diseases]. *Retrieved October 13, 2012, from http://www.vsn.nl/spierziekten*.

[602] F. Véricourt and O.B. Jennings. Nurse staffing in medical units: a queueing perspective. *Operations Research*, 59(6):1320–1331, 2011.

[603] I.B. Vermeulen, S.M. Bohte, S.G. Elkhuizen, P.J.M. Bakker, and H.L. La Poutré. Decentralized online scheduling of combination-appointments in hospitals. In *Proceedings of the International Conference on Automated Planning and Scheduling*, pages 372–379, 2008.

[604] I.B. Vermeulen, S.M. Bohte, S.G. Elkhuizen, J.S. Lameris, P.J.M. Bakker, and H.L. Poutré. Adaptive resource allocation for efficient patient scheduling. *Artificial Intelligence in Medicine*, 46(1):67–80, 2009.

[605] S. Villa, M. Barbieri, and F. Lega. Restructuring patient flow logistics around patient care needs: implications and practicalities from three critical cases. *Health Care Management Science*, 12(2):155–165, 2009.

[606] J.M.H. Vissers. Patient flow-based allocation of inpatient resources: a case study. *European Journal of Operational Research*, 105(2):356–370, 1998.

[607] J.M.H. Vissers, I.J.B.F. Adan, and N.P. Dellaert. Developing a platform for comparison of hospital admission systems: an illustration. *European Journal of Operational Research*, 180(3):1290–1301, 2007.

[608] J.M.H. Vissers and R. Beech, editors. *Health operations management: patient flow logistics in health care*, volume 2 of *Routledge Health Management Series*. Routledge, New York, NY, USA, 2005.

[609] J.M.H. Vissers and J. Wijngaard. The outpatient appointment system: design of a simulation study. *European Journal of Operational Research*, 3(6):459–463, 1979.

[610] N. Viswanadham and N.R.S. Raghavan. Performance analysis and design of supply chains: a Petri net approach. *Journal of the Operational Research Society*, 51(10):1158–1169, 2000.

[611] T.E. Vollmann, W.L. Berry, and C.D. Whybark. *Manufacturing planning and control systems*. McGraw-Hill, New York, NY, USA, 4th edition, 1997.

[612] M. Von Korff. A statistical model of the duration of mental hospitalization: the mixed exponential distribution. *Journal of Mathematical Sociology*, 6(2):169–175, 1979.

[613] R.E. Wachtel and F. Dexter. Tactical increases in operating room block time for capacity planning should not be based on utilization. *Anesthesia & Analgesia*, 106(1):215, 2008.

[614] D.T. Wade and B.A. De Jong. Recent advances in rehabilitation. *British Medical Journal*, 320(7246):1385–1388, 2000.

[615] J. Walrand. A note on Norton's theorem for queuing networks. *Journal of Applied Probability*, 20(2):442–444, 1983.

[616] J. Walrand. *An introduction to queueing networks*. Prentice Hall, Englewood Cliffs, NJ, USA, 1988.

[617] J. Walrand and P. Varaiya. Interconnections of Markov chains and quasi-reversible queuing networks. *Stochastic Processes and their Applications*, 10(2):209–219, 1980.

[618] L.M. Walts and A.S. Kapadia. Patient classification system: an optimization approach. *Health Care Management Review*, 21(4):75, 1996.

[619] J. Wang, S. Quan, J. Li, and A. Hollis. Modeling and analysis of work flow and staffing level in a computed tomography division of University of Wisconsin Medical Foundation. *Health Care Management Science*, 15(2):108–120, 2012.

[620] P.P. Wang. Sequencing and scheduling N customers for a stochastic server. *European Journal of Operational Research*, 119(3):729–738, 1999.

[621] W.Y. Wang and D. Gupta. Adaptive appointment systems with patient preferences. *Manufacturing and Service Operations Management*, 13(3):373–389, 2011.

[622] E.N. Weiss. Models for determining estimated start times and case orderings in hospital operating rooms. *IIE transactions*, 22(2):143–150, 1990.

[623] E.N. Weiss and J.O. McClain. Administrative days in acute care facilities: a queueing-analytic approach. *Operations Research*, 35(1):35–44, 1987.

[624] J.D. Welch and N.T.J. Bailey. Appointment systems in hospital outpatient departments. *The Lancet*, 259(6718):1105–1108, 1952.

[625] P.D. Welch. On the problem of the initial transient in steady-state simulation. *IBM Watson Research Center*, 1981.

[626] 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.

[627] G.P. Westert, J.S. Burgers, and H. Verkleij. The Netherlands: regulated competition behind the dykes? *British Medical Journal*, 339(7725):839–842, 2009.

[628] F. Wharton. On the risk of premature transfer from coronary care units. *Omega*, 24(4):413–423, 1996.

[629] M.J.B. White and M.C. Pike. Appointment systems in out-patients' clinics and the effect of patients' unpunctuality. *Medical Care*, 2(3):133–145, 1964.

[630] P. Whittle. Nonlinear migration processes. *Bulletin of the International Institute of Statistics*, 42:642–647, 1967.

[631] P. Whittle. Equilibrium distributions for an open migration process. *Journal of Applied Probability*, 5(3):567–571, 1968.

[632] P. Whittle. *Systems in stochastic equilibrium*. John Wiley & Sons, New York, NY, USA, 1986.

[633] P. Williams, G. Tai, and Y. Lei. Simulation based analysis of patient arrival to health care systems and evaluation of an operations improvement scheme. *Annals of Operations Research*, pages 1–17, 2010.

[634] S.V. Williams. How many intensive care beds are enough? *Critical Care Medicine*, 11(6):412, 1983.

[635] T. Williams and G. Leslie. Delayed discharges from an adult intensive care unit. *Australian Health Review*, 28(1):87–96, 2004.

[636] S. Winch and A.J. Henderson. Making cars and making health care: a critical review. *Medical Journal of Australia*, 191(1):28–29, 2010.

[637] W.L. Winston. *Operations research: applications and algorithms*. Duxbury Press, Boston, MA, USA, 2003.

[638] R.W. Wolff. *Stochastic modeling and the theory of queues*. Prentice Hall, Englewoood Cliffs, NJ, USA, 1989.

[639] World Health Organization. The world health report – Health systems financing: the path to universal coverage. *Retrieved October 13, 2012, from http://www.who.int/whr/2010/en/index.html*, 2010.

[640] World Health Organization & The World Bank. World report on disability. *Retreived October 13, 2012, from http://www.who.int/disabilities/en/*, 2011.

[641] D.J. Worthington. Queueing models for hospital waiting lists. *Journal of the Operational Research Society*, 38(5):413–422, 1987.

[642] D.J. Worthington. Hospital waiting list management models. *Journal of the Operational Research Society*, 42(10):833–843, 1991.

[643] M.B. Wright. The application of a surgical bed simulation model. *European Journal of Operational Research*, 32(1):26–32, 1987.

[644] P.D. Wright, K.M. Bretthauer, and M.J. Côté. Reexamining the Nurse Scheduling Problem: Staffing Ratios and Nursing Shortages. *Decision Sciences*, 37(1):39–70, 2006.

[645] C.H. Wu and K.P. Hwang. Using a discrete-event simulation to balance ambulance availability and demand in static deployment systems. *Academic Emergency Medicine*, 16(12):1359–1366, 2009.

[646] H. Xie, T.J. Chaussalet, and P.H. Millard. A continuous time Markov model for the length of stay of elderly people in institutional long-term care. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1):51–61, 2005.

[647] H. Xie, T.J. Chaussalet, and P.H. Millard. A model-based approach to the analysis of patterns of length of stay in institutional long-term care. *IEEE Transactions on Information Technology in Biomedicine*, 10(3):512–518, 2006.

[648] H. Xie, T.J. Chaussalet, W.A. Thompson, and P.H. Millard. A simple graphical decision aid for the placement of elderly people in long-term care. *Journal of the Operational Research Society*, 58(4):446–453, 2006.

[649] H.H. Xiong, M.C. Zhou, and C.N. Manikopoulos. Modeling and performance analysis of medical services systems using Petri nets. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, volume 3, pages 2339–2342, 1994.

[650] N. Yankovic and L.V. Green. Identifying good nursing levels: a queuing approach. *Operations research*, 59(4):942–955, 2011.

[651] S. Zeltyn, Y.N. Marmor, A. Mandelbaum, B. Carmeli, O. Greenshpan, Y. Mesika, S. Wasserkrug, P. Vortman, A. Shtub, and T. Lauterman. Simulation-based models of emergency departments: operational, tactical, and strategic staffing. *ACM Transactions on Modeling and Computer Simulation*, 21(4):24, 2011.

[652] B. Zhang, P. Murali, M.M. Dessouky, and D. Belson. A mixed integer programming approach for allocating operating room capacity. *Journal of the Operational Research Society*, 60(5):663–673, 2009.

[653] J. Zhang, J.E. Mason, B.T. Denton, and W.P. Pierskalla. Applications of operations research to the prevention, detection, and treatment of disease. In S. Gass and M. Fu, editors, *Encyclopedia of Operations Research and Management Science*, volume 1. Springer, New York, NY, USA, 3rd edition, 2011.

[654] W.H.M. Zijm. Towards intelligent manufacturing planning and control systems. *OR Spectrum*, 22(3):313–345, 2000.

[655] M.E. Zonderland, F. Boer, R.J. Boucherie, A. de Roode, and J.W. van Kleef. Redesign of a university hospital preanesthesia evaluation clinic using a queuing theory approach. *Anesthesia & Analgesia*, 109(5):1612–1614, 2009.

[656] M.E. Zonderland, R.J. Boucherie, N. Litvak, and C.L.A.M. Vleggert-Lankamp. Planning and scheduling of semi-urgent surgeries. *Health Care Management Science*, 13(3):256–267, 2010.

[657] R. Zurawski and M.C. Zhou. Petri nets and industrial applications: a tutorial. *IEEE Transactions on Industrial Electronics*, 41(6):567–583, 1994.

Acronyms

AAC	Acute Admission Cycle
AMC	Academic Medical Center Amsterdam
CAS	Cyclic Appointment Schedule
CHOIR	Center for Healthcare Operations Improvement and Research
CMCA	Children's Muscle Center Amsterdam
CPB	Netherlands Bureau for Economic Policy Analysis
CT	Computed Tomography
DSS	Decision Support System
ECG	Electrocardiogram
ED	Emergency Department
EMG	Electromyography
FCFS	First-Come First-Served
FTE	Full Time Equivalent
GDP	Gross Domestic Product
GEN	General Surgery
GLB	Group-Local-Balance
ICU	Intensive Care Unit
IFC	Inpatient Facility Cycle
ILP	Integer Linear Program
IOM	Institute of Medicine
KI	Kinesiologist
MAU	Medical Assessment Unit
MCU	Medium Care Unit
MeSH	Medical Subject Headings
MP	Myopathy
MRI	Magnetic Resonance Imaging
MSS	Master Surgical Schedule
MTM	Multidisciplinary Team Meeting
NMD	Neuromotor Disease
NP	Neuropathy

OECD	Organisation of Economic Co-operation and Development
OR/MS	Operations Research and Management Sciences
ORT	Orthopaedics
OT	Occupational Therapist
PACU	Post Anaesthesia Care Unit
PLA	Plastic Surgery
PS	Psychologist
PT	Physiotherapist
RP	Rehabilitation Physician
SMA	Spinal Muscular Atrophy
SPN	Stochastic Petri Net
SW	Social Worker
TRA	Traumatology
URO	Urology
VAS	Vascular Surgery
WHO	World Health Organization
WoS	Web of Science

Summary

During the upcoming decades, healthcare organizations face the challenge to deliver more patient care, of higher quality, and with less financial and human resources. The goal of this thesis is to help and guide healthcare professionals making their organizations future-proof. Building on techniques from Operations Research, a subfield of applied mathematics and economics, and focusing on the management of operations, the research presented contributes to a better understanding and functioning of healthcare delivery. The outcomes support decision makers in realizing the best possible use of available resources.

Demand for and expenditures on healthcare increase steadily, as a results of ageing populations, technological developments, and increased medical knowledge. At the same time, patient expectations, competition between healthcare organizations, and labor shortages are rising. With current efficiency levels being insufficient to keep healthcare affordable and accessible, let alone to be able to increase its quality, a joint effort is required by policy-makers, insurers, and care providers to fundamentally reconsider the way healthcare is delivered. This thesis is directed to the level of the healthcare providers, who are responsible for decisions about clinical practice and the management of healthcare operations.

The work presented intends to make healthcare professionals more aware of the added value of taking an integral perspective on logistical decision making. First, the problems addressed emphasize the importance of integrality in terms of objectives and performance: healthcare must be safe, effective, patient-centered, timely, efficient, and equitable. While the traditional belief is that quality and efficiency always confront each other, we demonstrate that they often can, and must, go hand in hand. Second, the research outcomes show the value of integrality in planning and control: performance is enhanced by aligning long-, medium-, and short-term decision making and by realizing coordination and collaboration between the various care chain actors. The results claim that taking an integral approach is the key to achieving what is reflected by the title of this dissertation: quality-driven efficiency.

The thesis is organized in six parts. Part I provides a general introduction. Part II provides an overview of the field of resource capacity planning and control in healthcare and a review of the state of the art in Operations Research. It sets up the conceptual framework within which several specific decision problems are studied in the following parts. Part III focuses on combination appointments during single outpatient visits, Part IV on multidisciplinary treatments requiring a series of outpatient visits, Part V on inpatient care services, and Part VI on entire care pathways.

Part I: Introduction

Within a healthcare organization, professionals of different disciplines jointly organize healthcare delivery. Designing and organizing processes is referred to by the term 'planning and control'; it involves setting goals and deciding in advance what to do, how to do it, when to do it and who should do it. Healthcare planning and control comprises multiple managerial functions, making medical, financial and resource decisions. This thesis addresses the managerial function of resource capacity planning and control, which concerns the dimensioning, planning, scheduling, monitoring, and control of renewable resources (i.e., facilities, equipment and staff).

The field of Operations Research and Management Sciences (OR/MS) is an interdisciplinary branch of applied mathematics and engineering that uses mathematical modeling to improve an organization's ability to enact rational and meaningful management decisions. The process of investigating a real-world problem via OR/MS starts with carefully observing and formulating the problem, including gathering all relevant data. The next step is to construct a mathematical model that abstracts the essence of the real problem. Next, by quantitatively predicting the consequences of potential interventions, the goal is to make recommendations to decision makers.

The research described is for a substantial part motivated by challenges faced in the organization of patient care at the Academic Medical Center (AMC) in Amsterdam, the Netherlands. These challenges are common to many present-day healthcare providers, and our mathematical models are generically formulated. Therefore, the application of the models and the relevance of their derived conclusions are not at all limited to the setting of the AMC. With the purpose to provide the best decision support in each particular problem setting, a diversity of OR/MS techniques (often in combination) is applied: computer simulation, heuristics, Markov processes, mathematical programming, queueing theory, and stochastic Petri nets.

Part II: A Taxonomy for Resource Capacity Planning and Control

This part comprises **Chapter 2** and provides a comprehensive overview of the typical decisions to be made in resource capacity planning and control in healthcare, in addition to a structured review of relevant OR/MS articles for each planning decision. Its contribution is twofold. First, to position the planning decisions, we present a taxonomy. This taxonomy provides healthcare managers and OR/MS researchers with a method to identify, break down and classify involved planning decisions. It contains two axes: the vertical axis reflects the hierarchical nature of decision making (strategic, tactical, and operational), and the horizontal axis the different healthcare services (ambulatory, emergency, surgical, inpatient, home, and residential care services). Second, following the taxonomy, for each of the services characterized, we provide an exhaustive specification of planning decisions. For each identified decision, we structurally review the key OR/MS literature and the OR/MS techniques that have been applied. With this conceptual framework, we aim to facilitate healthcare professionals in realizing comprehensive and cohorent decision making, and to provide researchers with a tool to formulate and position future research topics.

Part III: Facilitating the One-Stop Shop Principle

This part presents two studies that have the purpose to support the realization of one-stop shopping at ambulatory care services. In many settings it is highly valuable to patients to offer the combination of consultations, diagnostics, and treatments during a single visit. By one-stop shopping the number of visits can be reduced, and treatments can earlier be commenced and better be coordinated. **Chapter 3** is directed to outpatient clinics and diagnostic facilities that facilitate walk-in service, to improve accessibility, to offer more freedom for patients to choose their preferred time and date of visit, and to allow patients to visit multiple care providers on one day. The chapter shows the advantages of offering combined walk-in and scheduled service. **Chapter 4** provides an example of how OR/MS can support focused care facilities that offer multidisciplinary care to patients with specific complex diseases. The example concerns the 'Children's Muscle Center Amsterdam', which was opened in 2011 by the AMC to diagnose and treat children with neuromuscular diseases. Through the establishment of the center, clinical alignment is improved and children will generally visit the hospital only once a year instead of four to ten times.

Part IV: Coordinating Multidisciplinary Treatments

This part is directed to rehabilitation care. Rehabilitation care is a treatment process that involves a series of treatments by therapists of various disciplines. These therapists may be affiliated with different departments and may use different planning horizons. This multidisciplinary nature of the rehabilitation process complicates planning and control. **Chapter 5** presents a methodology to schedule treatments for rehabilitation outpatients entirely at once. This integral treatment planning methodology ensures continuity of the rehabilitation process while improving performance on various indicators among which access times, therapist utilization, and the ability to schedule combination appointments. The approach is applied to the rehabilitation outpatient clinic of the AMC. **Chapter 6** connects with the observation made at the end of Chapter 5, which states that balancing discipline capacities is a promising direction for further improvement. We perform an integral patient flow analysis for a case study of the rehabilitation center 'Het Roessingh', to support the implementation of treatment plans that are similar to those of Chapter 5. The general conclusion of Part IV is that facilitating coordination and alignment between different disciplines yields considerable improvements in both quality and efficiency.

Part V: Integrally Shaping Inpatient Care Services

This part aims to support the design and operations of inpatient care services. Effectively designing inpatient care services requires simultaneous consideration of several interrelated planning issues, such as case mix, care unit partitioning, care unit size, and staffing decisions. The inpatient care facility is a downstream department of which the workload is mainly determined by the patient outflow of the operating theater and the emergency department. Therefore, coordination with surgical and

emergency care services is essential. Workload on nursing wards depends highly on patient arrivals and patient lengths of stay, which are both inherently variable. Predicting this workload, and staffing nurses accordingly, is essential for guaranteeing quality of care in a cost effective manner. **Chapter 7** presents a model to predict bed census on nursing wards by hour as a function of the operating room schedule and the cyclic arrival pattern of emergency patients. The model enables the evaluation of alternative interventions with respect to both the design and the operations of inpatient care units. **Chapter 8** introduces a method which takes the hourly census predictions as starting point to derive efficient nurse staffing policies. In particular, it explores the potential of flexible staffing policies which allow hospitals to dynamically respond to their fluctuating patient population by employing float nurses. The effectiveness of both models is demonstrated by applying them to a case study of four surgical nursing wards of the AMC. The numerical results for this case study show that while the quality of delivered services becomes more reliable, the productivity of the beds and staff can be increased by roughly 10–20%. Inspired by these results, the AMC decided that the presented methods will be used during the upcoming years in supporting a complete redesign of the inpatient care facility.

Part VI: Modeling Care Chains with Stochastic Petri Nets

This part intends to model entire patient care pathways. These pathways are generally stochastic and various patient flows share different resources. Typical questions arising when designing healthcare organizations are the identification of bottlenecks, achievable throughput and maximization of resource utilization. Therefore, performance analysis is an important issue in the design and implementation of healthcare systems. We argue that stochastic Petri nets are an appropriate formalism to model interacting care pathways in healthcare organizations. We build a theoretical foundation for a decision support tool along which we believe vital insight in the behavior of healthcare networks can be obtained. **Chapter 9** serves as an introduction to the chapters that follow by outlining elementary Petri nets definitions, properties, and results, and by providing a review of relevant literature. **Chapter 10** focuses on analytical (so-called product form) results, to create the conditions for efficient computation of relevant performance measures via closed-form expressions. **Chapters 11 and 12** formulate decomposition results that contribute to greater understanding of network behavior and performance, as they enable studying a system by analyzing the characteristics of separate components. **Chapter 13** takes the described results as starting point, to sketch directions for future research aimed at constructing and evaluating stochastic Petri nets based on patient event logs, thereby becoming able to deliver practical decision support.

Conclusions

Planning and control has a rich tradition in manufacturing. The nature of healthcare operations inhibits direct copying of successful industry practices, as it has certain distinctive characteristics. Variability is a concept inherently attached to healthcare

operations, for example due to uncertainty of demand volumes and heterogeneity of patient's conditions and personalities. It complicates resource capacity planning and control, because standardization of operations is only desirable to a limited extent. The studies described effectively address the challenge of reducing artificial (created by deficiencies in planning and control) variability and anticipating natural (unavoidable, or even desirable) variability. Incorporating flexibility in planning creates the ability to specify and adjust planning decisions closer to the time of actual healthcare delivery. As a result, we show that it ensures a better match between care supply and fluctuating demand.

The value of creating clinical and logistical synergy is underlined by many of the chapters in this dissertation. In the first place, we demonstrate that realizing high-quality care delivery demands coordinated strategic, tactical and operational decision making. Recognizing and incorporating these hierarchical relations in decision making improves healthcare delivery performance. Second, since the clinical course is typically a highly fragmented process, facilitating coordination and collaboration between the actors within a care chain is shown to reduce clinical and logistical misalignment. This has positive consequences on patient outcomes, patient satisfaction, and resource utilization.

The value of applying Operations Research to healthcare delivery problems has been expressed in both its process and its outcomes. The process of modeling leads to better understanding and recognition of a problem. The outcomes of mathematical models make it possible to prospectively assess the consequences of various alternative interventions, without actually changing the system. Modeling is highly suitable in healthcare settings, since experimenting in practice may induce risks for patients and field experimenting takes more time, is more costly, and offers less statistical reliability. Moreover, since healthcare environments are generally politically charged, by quantifying the impact of potential solutions fact-based rather than feeling-based decision making can be realized.

In conclusion, this thesis demonstrates that Operations Research can play an essential role in addressing the tough logistical challenges healthcare organizations face. Mathematical modeling can make a positive contribution to the achievement of higher quality and increased productivity of labor and capital. We are convinced that healthcare organizations can benefit from giving mathematical modeling a permanent position in their decision-making processes. Because implementation of solutions often requires people to do things differently, it often meets with resistance. A prerequisite for successful implementation of results is that of operations researchers and practitioners working closely together. This thesis intends to build a bridge between science and practice.

Samenvatting

De gezondheidszorg staat de komende decennia voor de uitdaging om meer en betere zorg te verlenen met minder personele en financiële middelen. Het doel van dit proefschrift is om zorgprofessionals richting te bieden bij het toekomstbestendig maken van hun organisaties. Het gepresenteerde onderzoek bouwt op technieken uit de mathematische besliskunde (Operations Research) en richt zich op het ontwerpen en organiseren van zorgprocessen. Door het creëren van een beter inzicht in het functioneren van patiëntenzorg en het aanreiken van verschillende wiskundige methodes, helpen de onderzoeksresultaten zorgaanbieders optimaal gebruik te maken van hun beschikbare middelen.

Als gevolg van vergrijzing, technologische ontwikkelingen en toenemende medische kennis nemen de vraag naar zorg en de daarmee gepaard gaande kosten gestaag toe. Tegelijkertijd stijgen de verwachtingen van patiënten, groeit de competitie tussen zorgleveranciers en ontstaan personeelstekorten. De huidige efficiëntieniveaus zijn onvoldoende om de gezondheidszorg betaalbaar en toegankelijk te houden, laat staan om kwaliteitsverbeteringen mogelijk te maken. Een gezamenlijke inspanning van politici, verzekeraars en zorgaanbieders is daarom noodzakelijk om de manier waarop zorg wordt georganiseerd fundamenteel te herzien. Dit proefschrift heeft betrekking op het niveau van de zorgaanbieders: de verantwoordelijken voor de medische en logistieke besluitvorming in de patiëntenzorg.

Het gepresenteerde werk beoogt zorgprofessionals meer bewust te maken van de toegevoegde waarde van het kiezen van een integrale benadering van zorglogistiek. De behandelde vraagstukken benadrukken ten eerste het belang van integraliteit in doelstellingen: gezondheidszorg dient veilig, effectief, patiëntgericht, tijdig, efficiënt en gelijkwaardig te zijn. Waar van oorsprong de overtuiging heerst dat kwaliteit en efficiëntie lijnrecht tegenover elkaar staan, demonstreren wij dat ze vaak samen kunnen, en moeten, gaan. Ten tweede tonen we de waarde van integraliteit in management en bestuur: het goed op elkaar afstemmen van lange, middellange en korte termijnplanning werkt prestatiebevorderend. Hetzelfde geldt voor het realiseren van een goede coördinatie tussen de verschillende actoren die betrokken zijn in een zorgtraject. De onderzoeksresultaten illustreren dat het faciliteren van integrale samenwerking en besluitvorming de sleutel vormt tot het bereiken van wat wordt weerspiegeld door de titel van dit proefschrift: kwaliteitsgedreven efficiëntie.

Het proefschrift bestaat uit zes onderdelen. Deel I geeft een algemene inleiding tot het onderzoeksonderwerp. Deel II introduceert een conceptueel raamwerk voor capaciteitsplanning in de gezondheidszorg en geeft een overzicht van de gerelateerde state-of-the-art in de operations research literatuur. Hiermee scheppen we

het kader waarbinnen in de latere hoofdstukken verschillende specifieke logistieke vraagstukken worden bestudeerd. Deel III richt zich op het mogelijk maken van combinatieafspraken in de ambulante zorg, Deel IV op multidisciplinaire behandelingen bestaande uit een serie van poliklinische bezoeken, Deel V op het organiseren van klinische zorg, en Deel VI op het modelleren van volledige zorgtrajecten.

I Introductie

Dit deel bestaat uit **Hoofdstuk 1** en vormt de introductie van dit proefschrift. In zorginstellingen organiseren professionals van verschillende disciplines gezamenlijk de patiëntenzorg. De benaming voor het ontwerpen en organiseren van processen is planning en besturing; het omvat het formuleren van doelstellingen en het vooraf beslissen wat te doen, hoe en wanneer, en wie het zal doen. Planning en besturing van zorgprocessen vereist de betrokkenheid van meerdere managementgebieden, die samen medische, financiële, en capaciteitsbeslissingen nemen. De focus van dit proefschrift is het functiegebied van capaciteitsplanning en -besturing: het dimensioneren, plannen, roosteren, monitoren en beheersen van personele, instrumentele en ruimtelijke middelen.

Het wetenschapsgebied Operations Research en Management Sciences (OR/MS) is een multidisciplinaire tak van toegepaste wiskunde en economie. Het maakt gebruik van wiskundige modellen om het vermogen van een organisatie om te komen tot rationele en doelmatige besluitvorming te verbeteren. Het proces van het onderzoeken van een praktijkprobleem met behulp van OR/MS begint bij het zorgvuldig observeren en formuleren van het vraagstuk, samen met het verzamelen van alle relevante data. De volgende stap is het opstellen van een wiskundig model dat in abstracte vorm de essentie van het echte probleem vangt. Daarna is het doel om de consequenties van mogelijke praktijkinterventies kwantitatief te voorspellen, aan de hand waarvan aanbevelingen worden gedaan aan de probleemeigenaren.

De onderzoeksonderwerpen zijn voor een belangrijk deel gemotiveerd door uitdagingen die het Academisch Medisch Centrum (AMC) in Amsterdam ervaart in het organiseren van patiëntenzorg. De toepassing van onze wiskundige modellen en de geldigheid van afgeleide conclusies zijn zeker niet beperkt tot de setting van het AMC, aangezien deze uitdagingen herkenbaar zijn voor veel hedendaagse zorgaanbieders en de modellen generiek geformuleerd zijn. Met het doel om tot de beste beslissingsondersteuning te komen in elke specifieke probleemsituatie, hebben we verscheidene OR/MS technieken (veelal in combinatie) toegepast: computersimulatie, heuristieken, Markov processen, mathematisch programmeren, wachtrijtheorie en stochastische Petri netten.

II Een Taxonomie voor Capaciteitsplanning en -besturing

Dit deel bevat **Hoofdstuk 2**. Het geeft een uitgebreid overzicht van de beslissingen die aan de orde zijn in de capaciteitsplanning en -besturing van patiëntenzorg, met daarnaast een gestructureerd literatuuronderzoek van relevante OR/MS artikelen voor elke planningsbeslissing. De wetenschappelijke bijdrage is tweeledig. Ten

eerste presenteren we een taxonomie ter positionering van de planningsbeslissingen. Dit reikt zorgmanagers en OR/MS onderzoekers een methode aan om planningsbeslissingen te kunnen identificeren en classificeren. De taxonomie bestaat uit twee assen: de verticale as weerspiegelt het hiërarchische karakter van besluitvorming (strategische, tactische en operationele planning), en de horizontale as de verschillende typen zorgverlening (ambulante, spoedeisende, chirurgische, klinische, thuis- en residentiële zorg). Ten tweede stellen we aan de hand van de taxonomie een uitvoerige specificatie van beslissingen op. Voor elke geïdentificeerde planningsbeslissing bestuderen we de belangrijkste OR/MS literatuur en de wiskundige technieken die daarin zijn toegepast. Met het presenteren van dit conceptuele raamwerk beogen we zorgprofessionals te faciliteren in het realiseren van volledige en coherente besluitvorming, en beogen we onderzoekers een instrument te bieden ter inspiratie voor en positionering van toekomstige onderzoeksonderwerpen.

III Het Faciliteren van het One-Stop Shop Principe

Dit deel presenteert twee studies die tot doel hebben de organisatie van ambulante zorg via het 'one-stop shop principe' te ondersteunen. In veel gevallen is het voor patiënten zeer waardevol om binnen één bezoek de benodigde consulten, diagnostische onderzoeken en behandelingen gecombineerd te kunnen ontvangen. Middels 'one-stop shopping' wordt het aantal bezoeken teruggebracht, en kunnen behandelingen eerder worden gestart en beter worden gecoördineerd. **Hoofdstuk 3** richt zich op poliklinieken en diagnostische faciliteiten die zorg op inloop aanbieden, om zo de toegankelijkheid te verbeteren, patiënten meer vrijheid te bieden om hun voorkeursdag en -tijd te kiezen en het bezoeken van meerdere zorgverleners op één dag mogelijk te maken. Dit hoofdstuk toont de voordelen van het slim combineren van een inloop- en een afspraaksysteem. **Hoofdstuk 4** geeft een voorbeeld van hoe OR/MS instellingen kan ondersteunen die zich exclusief richten op zorgverlening aan patiënten met specifieke complexe aandoeningen. Het betreft het 'Kinderspiercentrum Amsterdam', dat in 2011 door het AMC is geopend om kinderen met een spierziekte te diagnosticeren en behandelen. Door het oprichten van dit centrum is de multidisciplinaire zorg beter op elkaar afgestemd en hoeven de meeste kinderen het ziekenhuis slechts eens per jaar te bezoeken in plaats van vier tot tien keer.

IV Het Coördineren van Multidisciplinaire Behandelingen

Dit deel richt zich op de organisatie van revalidatiezorg. Een behandelproces van een revalidatiepatiënt bestaat doorgaans uit een serie behandelingen uitgevoerd door therapeuten van verschillende disciplines. Vaak werken deze therapeuten voor verschillende afdelingen en hanteren zij een verschillende planningshorizon. Dit multidisciplinaire karakter bemoeilijkt de planning en besturing van revalidatiezorg. **Hoofdstuk 5** presenteert een algoritme waarmee volledige behandeltrajecten voor poliklinische patiënten in één keer gepland kunnen worden. Het toepassen van deze integrale planningsmethodiek verzekert de continuïteit van het behandelproces en verbetert prestatie-indicatoren als toegangstijd en bezettingsgraad van therapeuten,

evenals het vermogen om combinatieafspraken te kunnen aanbieden. Deze methodiek wordt toegepast op de Polikliniek Revalidatie van het AMC. **Hoofdstuk 6** sluit aan op de observatie van het voorgaande hoofdstuk dat door het balanceren van de capaciteiten van verschillende disciplines verdere verbeteringen mogelijk zijn. We maken een integrale analyse van de patiëntstromen voor een case study van het revalidatiecentrum 'Het Roessingh' om ondersteuning te bieden voor de geplande implementatie van behandelplannen die vergelijkbaar zijn met die uit Hoofdstuk 5. De algemene conclusie van Deel IV is dat het bevorderen van coördinatie en afstemming tussen disciplines verbeteringen oplevert in zowel kwaliteit als efficiëntie.

V Het Integraal Vormgeven van Klinische Zorgprocessen

Dit deel richt zich op beslissingsondersteuning voor klinische zorg. Het effectief organiseren van klinische zorg vraagt het in samenhang nemen van een reeks verweven planningsbeslissingen (zoals het bepalen van de case mix, de indeling en grootte van verpleegafdelingen, en de personeelsplanning). De werklast op verpleegafdelingen hangt in hoge mate samen met de inherent variabele patiëntvolumes en ligduren. Het voorspellen van deze werklast en het hierop laten aansluiten van de personeelsroosters is essentieel om kwalitatief hoogwaardige zorg te kunnen leveren tegen acceptabele kosten. Klinische zorg is onderdeel van een zorgketen: de instroom van patiënten op verpleegafdelingen wordt voornamelijk bepaald door de uitstroom van de operatiekamers en de spoedeisende hulp. Daarom is afstemming met deze twee afdelingen zeer wenselijk. **Hoofdstuk 7** presenteert een model om de bedbezetting te voorspellen op uurniveau als functie van het operatieschema en het cyclische aankomstpatroon van spoedpatiënten. Met het model kunnen de consequenties van alternatieve interventies met betrekking tot het ontwerp en de organisatie van klinische zorgprocessen doorgerekend worden. **Hoofdstuk 8** introduceert een model dat de uurlijkse bedbezettingvoorspellingen als uitgangspunt neemt om efficiënte verpleegkundige inzet te bepalen. Het verkent in het bijzonder het potentieel van het inzetten van flexibele verpleegkundigen. Door pas aan het begin van een dienst te bepalen op welke verpleegafdeling zo'n flexibele verpleegkundige werkt, zijn ziekenhuizen in staat om dynamisch te reageren op hun fluctuerende patiëntenpopulatie. De effectiviteit van beide methodes wordt gedemonstreerd aan de hand van een case study betreffende vier chirurgische verpleegafdelingen van het AMC. De numerieke resultaten voor deze case laten zien dat de productiviteit van personeel en bedden met 10–20% kan worden verbeterd, terwijl de kwaliteit van de geleverde zorg betrouwbaarder wordt. Geïnspireerd door deze resultaten heeft het AMC besloten dat de methodes de komende jaren gebruikt gaan worden ter ondersteuning van een compleet herontwerp van de klinische zorg.

VI Het Modelleren van Zorgketens met Stochastische Petri Netten

Dit deel richt zich op het modelleren van volledige zorgtrajecten. In het algemeen zijn deze trajecten stochastisch en worden capaciteiten gedeeld door verschillende patiëntstromen. Karakteristieke vraagstukken bij het ontwerpen van zorgprocessen

zijn het identificeren van knelpunten in patiëntstromen, het vaststellen van haalbare patiëntvolumes en het maximaliseren van bezettingsgraden. Prestatieanalyse kan daarom een wezenlijke bijdrage leveren aan het ontwerpen en stroomlijnen van zorgprocessen. We beargumenteren dat stochastische Petri netten een geschikt formalisme vormen om interacterende patiëntstromen te modelleren. We leggen een theoretische basis voor een beslissingsondersteunend systeem (decision support system) aan de hand waarvan inzicht in het gedrag van zorgnetwerken kan worden verkregen. **Hoofdstuk 9** dient als introductie op de daaropvolgende hoofdstukken door het beschrijven van elementaire Petri net definities, eigenschappen en resultaten, en door het geven van een overzicht van relevante literatuur. **Hoofdstuk 10** concentreert zich op analytische (zogenaamde produktvorm) resultaten, om zo de voorwaarden te scheppen voor efficiënte doorrekenng van relevante prestatiematen. **Hoofdstukken 11 en 12** formuleren decompositieresultaten die bijdragen tot een beter begrip van het gedrag en de prestatie van een netwerk, doordat ze het mogelijk maken om een systeem te bestuderen aan de hand van de karakteristieken van afzonderlijke componenten. **Hoofdstuk 13** neemt de beschreven resultaten als startpunt om richtingen voor toekomstig onderzoek te schetsen dat gericht is op het construeren en evalueren van stochastische Petri netten gebaseerd op patient event logs (digitaal opgeslagen gebeurtenissen), om daarmee daadwerkelijke praktische beslissingsondersteuning te kunnen gaan bieden.

Conclusies

In de maakindustrie heeft capaciteitsplanning en -besturing een rijke traditie. Het specifieke karakter van de zorgsector maakt dat het niet mogelijk is om succesvolle concepten uit de industrie direct te kopiëren. Variabiliteit is één van de onderscheidende kenmerken die onlosmakelijk verbonden zijn met zorgverlening, voortkomend bijvoorbeeld uit onzekerheid in vraagvolumes en heterogeniteit in ziektebeelden en persoonlijkheden van patiënten. Het compliceert capaciteitsplanning, omdat hierdoor standaardisatie van zorgprocessen slechts in beperkte mate wenselijk is. De beschreven studies adresseren de uitdaging van het reduceren van kunstmatige variatie (gecreëerd door onregelmatigheden in capaciteitsplanning) en het anticiperen op natuurlijke (onvermijdelijke, of zelfs onwenselijke) variatie. Het inbouwen van flexibiliteit schept de mogelijkheid om planningsbeslissingen dichter op het moment van de daadwerkelijke zorgverlening te specificeren. Wij laten zien dat flexibiliteit de mogelijkheid biedt de afstemming van het zorgaanbod op fluctuerende vraag te verbeteren.

De waarde van het creëren van klinische en logistieke synergie wordt onderstreept door dit proefschrift. In de eerste plaats demonstreren we dat het realiseren van hoogwaardige zorg nauwkeurig op elkaar afgestemde strategische, tactische en operationele besluitvorming vereist. Het onderkennen van de hiërarchische relaties tussen beslissingen verbetert de prestaties van de zorgverlening. In de tweede plaats wordt aangetoond dat, doordat de meeste zorgtrajecten zeer gefragmenteerd zijn, het faciliteren van samenwerking tussen de actoren in een zorgketen klinische en logistieke coördinatie verbetert. Dit heeft een positieve uitwerking op patiënt-

uitkomsten, patiënttevredenheid en efficiënt gebruik van middelen.

De waarde van het toepassen van Operations Research op vraagstukken in de organisatie van gezondheidszorg is tot uitdrukking gekomen zowel in het proces als de uitkomsten van de modelleerexercities. Het proces van modelleren leidt tot betere erkennning en gezamenlijk begrip van voorliggende vraagstukken. De uitkomsten van wiskundige modellen maken het mogelijk om de effecten van potentiële oplossingen prospectief te kwantificeren en daarmee objectief in te schatten zonder dat er hoeft te worden ingegrepen in de praktijk. Modelleren is met name geschikt in zorgomgevingen omdat veldexperimenten risico's voor patiënten met zich meebrengen, meer tijd en geld kosten, en minder statistische betrouwbaarheid opleveren. Omdat besluitvormingsprocessen in de zorg vaak politiek gevoelig zijn, kan het kwantificeren van de impact van mogelijke keuzes er bovendien toe leiden dat besturing van een organisatie minder op gevoel en meer op feiten gebaseerd wordt.

Op basis van dit proefschrift kan geconcludeerd worden dat Operations Research een essentiële rol kan spelen in het aanpakken van de lastige uitdagingen waar de gezondheidszorg voor staat. Wiskundig modelleren levert een positieve bijdrage aan het bereiken van betere kwaliteit en hogere productiviteit van arbeid en kapitaal. Wij zijn ervan overtuigd dat zorginstellingen er baat bij hebben om wiskundig modelleren een vaste plek te geven in hun besluitvormingsprocessen. Implementatie van oplossingen gaat vaak gepaard met veranderingen in het dagelijks functioneren van mensen. Een nauwe samenwerking tussen onderzoekers en zorgprofessions is daarom een randvoorwaarde voor succesvolle implementatie. Dit proefschrift beoogt een verbinding te leggen tussen wetenschap en prakijk.

About the author

Nikky Kortbeek was born in Beverwijk, the Netherlands, on November 1, 1983. In 2001, he obtained his gymnasium degree at Gymnasium Felisenum in Velsen-Zuid, after which he commenced his studies at the University of Amsterdam (UvA). His combined interest in science and society led him to initially enroll in the 'beta-gamma propedeuse', a one-year program covering a range of technical and social studies. After having finished the beta-gamma propedeuse, and a propedeuse in Psychology, he obtained a cum laude Bachelor's of Science degree in Econometrics and Operations Research & Management in 2006, with a thesis on dimensioning intensive care units. He graduated in 2008 with a cum laude Master's of Science degree in Operation Research & Management, with a thesis on inventory management of blood platelets.

After being a research fellow for one year at the UvA, he joined the department Applied Mathematics of the University of Twente (UT), for a Ph.D. program with the Stochastic Operations Research group. He combined doing research with being a consultant patient logistics at the department of Quality and Process Innovation of the Academic Medical Center (AMC) in Amsterdam. In 2012, during a three-month research visit to Australia, he worked at the University of Melbourne, the University of Western Sydney, and Campbelltown Hospital. His Ph.D. research culminates with this dissertation.

The next step in Nikky's professional life will again be on the interface between science and practice. At the AMC, he will be appointed as a process consultant and as research program leader in healthcare logistics. In addition, he will be appointed a postdoctoral position at the UT research group CHOIR (Center for Healthcare Operations Improvement and Research) in the area of healthcare logistics.

List of publications

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

N. Kortbeek and R.J. Boucherie. A *P*- and *T*-invariant characterization of product form and decomposition in stochastic Petri nets. *Performance Evaluation*, 69(11): 573–599, 2012.
(Basis for Chapters 10 and 11).

N. Kortbeek, A. Braaksma, C.A.J. Burger, R.J. Boucherie, and P.J.M. Bakker. *Flexible nurse staffing based on hourly bed census predictions*. Memorandum 1996, Department of Applied Mathematics, University of Twente, Enschede, the Netherlands, 2012.
(Basis for Chapter 8).

N. Kortbeek, R.J. Boucherie, E. van Ommeren, and P.G. Taylor. *Structural characterization of decomposition in rate-insensitive stochastic Petri nets*. Memorandum 1993, Department of Applied Mathematics, University of Twente, Enschede, the Netherlands, 2012.
(Basis for Chapter 12).

M.F. van der Velde, N. Kortbeek, and N. Litvak. *Organizing multidisciplinary care for children with neuromuscular diseases*. Memorandum 1991, Department of Applied Mathematics, University of Twente, Enschede, the Netherlands, 2012.
(Basis for Chapter 4).

N. Kortbeek, A. Braaksma, H.F. Smeenk, P.J.M. Bakker, and R.J. Boucherie. *Integral resource capacity planning for inpatient care services based on hourly bed census predictions*. Memorandum 1990, Department of Applied Mathematics, University of Twente, Enschede, the Netherlands, 2012.
(Basis for Chapter 7).

N. Baer, N. Kortbeek, N. Litvak, and O. Roukens. *Patient flow analysis in pain rehabilitation care*. Memorandum 1989, Department of Applied Mathematics, University of Twente, Enschede, the Netherlands, 2012.
(Basis for Chapter 6).

A. Braaksma, N. Kortbeek, G.F. Post, and F. Nollet. *Integral multidisciplinary rehabilitation treatment planning*. Memorandum 1985, Department of Applied Mathematics, University of Twente, Enschede, the Netherlands, 2012.
(Basis for Chapter 5).

A. Fügener, E.W. Hans, R. Kolish, N. Kortbeek, and P.T. Vanberkel. *Master surgery scheduling with consideration of multiple downstream units*. Under review with European Journal of Operational Research, 2012.

N. Kortbeek, M.E. Zonderland, R.J. Boucherie, N. Litvak, and E.W. Hans. *Designing cyclic appointment schedules for outpatient clinics with scheduled and unscheduled patient arrivals*. Memorandum 1968, Department of Applied Mathematics, University of Twente, Enschede, the Netherlands, 2011.
(Basis for Chapter 3).

P.J.H. Hulshof, R.J. Boucherie, J.T. van Essen, E.W. Hans, J.L. Hurink, N. Kortbeek, N. Litvak, P.T. Vanberkel, E. van der Veen, B. Veltman, I.M.H. Vliegen, and M.E. Zonderland. ORchestra: an online reference database of OR/MS literature in health care. *Health Care Management Science*, 14(4):383–384, 2011.

W.L.A.M. de Kort, M. Janssen, N. Kortbeek, N. Jansen, J. van der Wal, and N.M. van Dijk. Platelet pool inventory management: theory meets practice. *Transfusion*, 51(11):2295–2303, 2011.

R. Haijema, N. Kortbeek, J. van der Wal, and N.M. van Dijk. Bloedstollende operations research. *StatOR* (Refereed journal of the Dutch association of Statistics and Operations Research), 11(2):23–26, 2010.

N. Kortbeek and N.M. van Dijk. *On the rejection probability in OT-ICU Systems*. AE Report 1/10, Faculty of Actuarial Science and Econometrics, University of Amsterdam, the Netherlands, 2010.

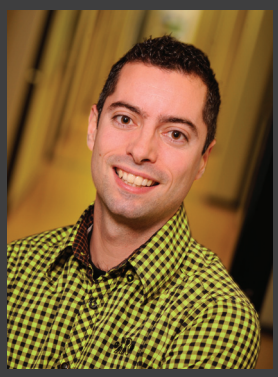
N.M. van Dijk and N. Kortbeek. Erlang loss bounds for OT-ICU systems. *Queueing Systems*, 63(1):253–280, 2009.

N. Kortbeek, J. van der Wal, N.M. van Dijk, R. Haijema, and W.L.A.M. de Kort. *Blood bank production and issuing optimization: strategies for younger platelets*. AE Report 3/08, Faculty of Actuarial Science and Econometrics, University of Amsterdam, the Netherlands, 2008.

During the upcoming decades, healthcare organizations face the challenge to deliver more patient care, of higher quality, and with less financial and human resources.

The goal of this thesis is to help and guide healthcare professionals making their organizations future-proof. Building on techniques from Operations Research, a subfield of applied mathematics, and focusing on the management of operations, the research presented contributes to a better understanding and functioning of healthcare delivery.

By creating clinical and logistical synergy, the outcomes support decision makers in realizing the best possible use of available resources.



ISBN. 978-90-365-3428-4

UNIVERSITY OF TWENTE.

Department of Applied Mathematics,
Stochastic Operations Research Group,
Center for Healthcare Operations Improvement and Research

amc

Beta
Research School for Operations Management and Logistics