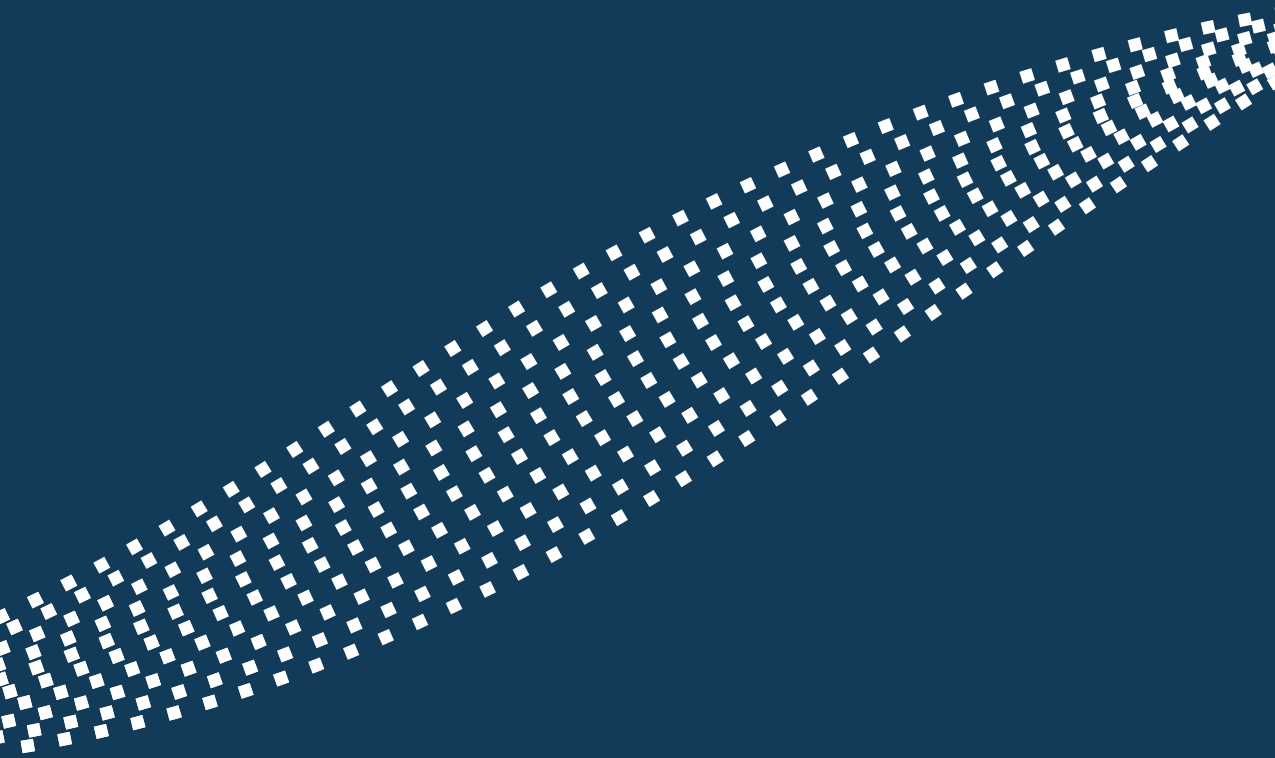
Beta

Research School for Operations Management and Logistics

Interacting Hospital Departments and Uncertain Patient Flows: Theoretical Models and Applications



Peter Tulkens Vanberkel

Interacting Hospital Departments and Uncertain Patient Flows: Theoretical Models and Applications

Peter Tulkens Vanberkel

Dissertation committee

Chairman & secretary	Prof. dr. P.J.J.M. van Loon
	Prof. dr. A.J. Mouthaan
Promotors	Prof. dr. R.J. Boucherie
	Prof. dr. J.L. Hurink
Assistant-promotor	Dr. ir. E.W. Hans
Members	Prof. dr. R. Kolisch
	Dr. J.T. Blake
	Prof. dr. M.J. IJzerman
	Prof. dr. W.H. van Harten
	Dr. N. Litvak

This thesis is number D144 of the thesis series of the Beta Research School for Operations Management and Logistics. The Beta Research School is a joint effort of the departments of Technology Management, and Mathematics and Computing Science at the Technische Universiteit Eindhoven and the Centre for Telematics and Information Technology at the University of Twente. Beta is the largest research centre in the Netherlands in the field of operations management in technology-intensive environments. The mission of Beta is to carry out fundamental and applied research on the analysis, design, and control of operational processes.

Ph.D. thesis, University of Twente, Enschede, the Netherlands
Center for Telematics and Information Technology (No. 11-198, ISSN 1381-3617)
Center for Healthcare Operations Improvement and Research

Printed by Ipskamp Drukkers BV, Enschede, the Netherlands

Cover design: The weave of white dots forming the asymmetric pattern abstractly represent an uncertain flow of patients. The weave culminates in the middle of multiple heterogeneous particles which together form the single image on the back cover. This image abstractly represents multiple departments interacting to form a single hospital. All design elements are part of the University of Twente's branding strategy.

isbn 978-90-365-3179-5

Interacting Hospital Departments and Uncertain Patient Flows: Theoretical Models and Applications

Dissertation

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
Prof. dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Friday, May 27th, 2011 at 14:45

by

Peter Tulkens Vanberkel,

born September 14th, 1981
in Antigonish, Canada.

This dissertation is approved by promotors,

Prof. dr. Richard J. Boucherie
Prof. dr. Johann L. Hurink

and assistant-promotor,
Dr. ir. Erwin W. Hans

For Connie and Caitlin

Contents

Chapter 1

Introduction

Contents

1.1 Challenges in health care delivery

Health care constitutes the largest industry in many developed countries [36], and managing it is a complex task due to its importance to society and the often politically charged atmosphere within which it exists. Furthermore, the nature of health care delivery does not allow the direct copying of success stories from the manufacturing industry, where logistical optimization has a long history. Health care processes and supply chains show considerable differences, such as a high degree of uncertainty, the medical autonomy of clinicians, and the fact that

patients cannot be treated as products. The evolution of management philosophies seen over the past decades in the manufacturing industry offers a glimpse into the changes required in health care delivery processes. This is most evident in the change by modern manufacturers from a reductionist approach to a systems approach to management.

The *reductionist approach* to management employs the principles of F.W. Taylor, which sees management decisions being made by separately analyzing component parts. Using this approach, manufacturers can improve overall operations by decomposing work into specific tasks and then improving the efficiency of these tasks. However, too much emphasis on individual tasks can cause a loss of perspective of the overall system. In contrast, the *systems approach* to management focuses less on the individual tasks and more on their interactions. In understanding how tasks affect each other, managers can create a seamless environment where the overall work is completed in an efficient manner. Finding the balance between these two approaches is a challenge for managers and operations researchers alike.

As an example, consider the time required to changeover a machine from producing one part to producing a different part (called setup time). Managers using a reductionist approach determine the time when the department should switch from producing one part to producing another part, such that setup costs and inventory costs are balanced (e.g. producing too many products before switching results in large and expensive inventories, whereas switching too often is inefficient since no parts are produced while switching). The resulting optimal "switching time policy" stipulates the production schedule for the department. However, optimizing the switching time may not be the best solution when adjacent departments are considered, as the switching time policy does not take into account the supply of raw materials from any "upstream" departments or the needs of any "downstream" departments. For example, producing parts according to an optimal switching time policy may result in the department producing part A when the downstream department needs part B. Managers using a systems approach develop production schedules which account for

the operations of the adjacent departments, although possibly at the expense of the optimal switching time policy.

Since the 1980s, there has been a shift in the manufacturing industry from predominantly using a reductionist approach to predominantly using a systems approach. This switch has resulted in (among other things) lower production costs and shorter production times. The contrast in these management styles and their advantages and disadvantages are discussed in [95]. In the health care industry however, the shift from a reductionist approach to a systems approach has been lagging [193].

In health care there are natural pressures that cause managers to lose sight of the overall perspective and take a reductionist (or individual component) approach. Often "management does not consider the total care chain from admission to discharge, but mainly focuses on the performance of individual departments. Not surprisingly, this has often resulted in diminished patient access without any significant reduction in costs" [54]. This is further complicated because an "individual component" in the health care context is a living and breathing patient.

As argued in Chapter 2, health care literature is rife with studies on scheduling, resource utilization, and patient flow. However, these studies are often confined to the operation of a single department and ignore many of the complex relationships that exist between departments. As an example, patient arrival patterns are modelled with statistical distributions instead of as a consequence of previous care. This disjointed approach fails to offer coordinated patient trajectories and essentially represents a hospital as a collection of processes receiving patients from, and feeding patients into, buffers. From industry, we have learned that disjointed and unbalanced production lines lead to high buffer volumes, much work in process, long product cycle times and are plagued with inefficiencies [95].

It is my experience that the impact of disjointed operations are particularly serious in health care settings. Waiting patients, unlike waiting products, may complain, be prioritized and reprioritized, require on-

going care, and cause other excessive coordination and management efforts. For inpatients, waiting costs are high and direct, making the reduction of the length of stay of inpatients a priority in hospitals and a common goal of many studies. For outpatients, the costs associated with waiting are not direct and often hidden. In addition to the administrative costs, the quality of life costs for waiting outpatients are substantial. Besides the obvious extended period of time in poor health, there is anxiety associated with waiting, the possibility of further health deterioration, the loss of confidence in the hospital or physician and the compounded effect of all of these factors together.

Some headway toward modelling hospitals as a complete system is evident in health care literature [75, 193]. Many models consider the impact of their operations on downstream inpatient wards. Typical examples include bed occupancy dictated by the operating room schedule, and emergency department congestion caused by an inability to admit patients to an already overcrowded ward. There is also literature concerning a hospital's inability to discharge patients into long-term care. Hospitals are developing ambulatory care centres that locate multiple specialties together so that a patient's ambulatory treatment can, at the least, happen in the same space, and at the best, be efficiently coordinated.

In Chapter 2 we present a review of models used to examine issues related to patient flows, and illustrate the extent to which models account for interactions between the main department under study and adjacent departments. The review found only 88 papers describing patient flow models that considered resources from multiple hospital departments. This amount is consistent with findings of other authors [74, 110] who conclude that although there is an abundance of models for health care processes, few consider multiple units or departments.

We conclude that researchers often model hospitals in a way that reflects the reductionist view of managers. Models often consider only a single department and overlook the complex relationships that exist between departments. We believe that this approach is in response to two adverse but common characteristics of health care. The first is

the complexity that is inherent in health care and the second is the uncertainty in patient flows. The work in this thesis contributes to health care logistics research by addressing a series of complex problems related to interacting hospital departments and uncertain patient flows.

1.2 Hierarchical decision making and operations research

Competitive manufacturing companies make planning and control decisions in a hierarchical manner [210]. Long term strategic decisions are made at the highest hierarchical level and decisions relating to specific issues in real-time are made at the lowest hierarchical level. For example, the decision of what products to manufacture is at the top of the hierarchy and the decision of whether to discard a specific part due to its quality is at the bottom of the hierarchy. It follows therefore that decisions at the bottom of the hierarchy are guided by decisions made at the top of the hierarchy. In general, this reliance of one decision on another defines the hierarchy. Many planning and control frameworks classify decisions into three hierarchical levels: 1) strategic 2) tactical and 3) operational (as suggested first in 1965 in [5]). In the field of health care, many similar hierarchical planning and control frameworks have been proposed (see [88]).

As an example, consider a series of decisions related to the installation of public heart defibrillators. Having heart defibrillators available in public places allows prompt application which dramatically improves the likelihood of survival (and neurological recovery) in the event of cardiac arrest [104]. Typical decisions are as follows:

Strategic decisions: Should heart defibrillators be placed in public places? How much money should be spent on heart defibrillators?

Tactical decisions: Once the decision to act has been made and the amount of money to be spent is known, the tactical decisions can be

addressed. Where should the heart defibrillators be installed to maximize the public benefit? How many should be installed in each city? At what locations within cities (e.g. train stations, sport facilities, retirement homes) should they be installed?

Operational decisions: Once the locations are known, operational decisions can be addressed. How high should they be attached to the wall? How should they be maintained? How can their locations be clearly indicated?

For each decision at each level there are tradeoffs which need to be considered before making a final decision. For example, at the strategic level there is a direct tradeoff between the amount of money spent and the expected number of lives to be saved. At the operational level, a tradeoff exists between attaching the machines high enough to be out of reach of children and low enough such that the majority of adults can reach them. Evaluating these tradeoffs in a scientific manner to support decision makers is one use of operations research models.

The decision of how high to attach the machine can be based on a rather straightforward model. The modeller essentially uses historical population statistics to determine the percentage of children that cannot reach a certain height and the percentage of adults that can. Using historical population statistics to model the future population is a relatively accurate (or concrete) way to represent the actual system. The model assumes only that the future population will be similar (in height) to the current population. To model other more complex systems (and decisions), requires more assumptions and a more abstract view.

For example, to determine the appropriate investment in heart defibrillators, the modeller needs to determine the expected number of lives to be saved for a given investment. For this, the modeller must make assumptions about the future prevalence of heart disease, the probability of cardiac arrest happening within the vicinity of a machine, the probability of someone finding and using a machine, etc. As such, this decision requires more assumptions and as a result, the model becomes

a more abstract representation of the actual system.

Leading thought on how to manage complex organizations suggests that decisions be made in a hierarchical manner. Making decisions however, requires evaluating the tradeoffs between two or more measures. Supporting decision makers to understand these tradeoffs and the implication of certain actions, is the scientific discipline of operations research. Using operations research models to support logistical decision making (at multiple hierarchical levels) within hospitals is the underlying theme of this thesis.

1.3 Thesis structure

The work of this thesis is organized according to the hierarchical level of the decision being addressed. Beginning in Chapter 3 (Chapter 2 is a survey of literature), the hierarchical relationship of these decisions is as follows. Chapter 3 addresses a strategic patient mix decision used to determine which patient types should be treated at a hospital to meet case mix and capacity restrictions. In Chapter 4, the patient mix is assumed to be known and we consider the decision of whether (and to what extent) to pool resources. In Chapter 5 we consider a specific case where the decision not to pool resources has been made and answer the question of how many patients a single unpooled oncology clinic can follow. In Chapter 6 the pooling and patient mix decisions are assumed, and we answer the question of how two departments which treat patients *consecutively* can do so in a seamless manner. In Chapter 7, we also assume the pooling and patient mix decisions are made and we answer the question of how two departments that treat patients *concurrently* can do so in a seamless manner.

Similarly, the models of the initial chapters (and in particular Chapters 3 and 4) are more abstract and encompass a larger portion of the hospital. Chapter 3 models all departments but represents these departments in an abstract manner (i.e. departments are modelled by their monthly treatment volumes). Chapters 4 and 5 consider only

	Chapter title	Departments	Approach
2	Survey of literature		
3	Patient mix optimization	All	Mathematical programming and queueing theory
4	Efficiency evaluation for pooling resources	Consultation departments	Queueing theory and simulation
5	Panel sizing in oncology	Consultation departments	Queueing theory
6	Surgical scheduling and inpatient wards	Operating room and inpatient wards	Applied probability
7	Pharmacy policies to reduce waiting times	Pharmacy and consultation departments	Queueing theory and simulation

Table 1.1 – Chapter scopes and approaches

consultation departments (outpatient clinics) but model them more explicitly by taking into account characteristics of the department, such as the amount of resources, and treatment capacity per day. The analysis and results are sufficiently general for any consultation department regardless of the treatment being provided. Chapters 6 and 7 model the interactions between specific departments. Chapter 6 models the relationship between inpatient wards and the operating room and Chapter 7 models the relationship between the chemotherapy department and the pharmacy.

The models employed and developed in this thesis relate in general to the fields of queueing theory, mathematical programming and simulation. Table 1.1 summarizes the departments to which each chapter relates and also the modelling technique that is used (this classification is consistent with the literature review of Chapter 2). Specifically, we use $G/G/c$ queueing models / Lindley's recursion in Chapters 4, 5 and 7, infinite server queueing models in Chapters 3, 5 and 6 and mathematical programming in Chapter 3.

The chapters in this thesis are written in a self contained manner.

Each chapter (other than the survey of literature chapter) contains the following four sections: 1) *Introduction* 2) *Model description* 3) *Application* and 4) *Discussion* and where appropriate, additional sections are used. In the *Introduction* the problem is introduced, literature specific to the problem is reviewed, and the chapter's goals and structure are stated. The contents of the *Model description* and *Application* sections are self explanatory. In the *Discussion* section, main results are summarized, and future research potential is discussed.

1.4 Applied research environment

The decisions modelled in this thesis are motivated by actual problems faced by two cancer hospitals. Chapters 3, 4, 6 and 7 were motivated by problems at the Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital (NCI). NCI is a comprehensive cancer centre, which provides hospital care and research, and is located in Amsterdam, the Netherlands. The hospital has 150 inpatient beds and the outpatient department receives approximately 24,000 appointment requests every year.

Chapter 5 is motivated by a problem at the British Columbia Cancer Agency (BCCA), Canada. BCCA operates five cancer hospitals providing diagnostic services, chemotherapy, radiation therapy, and supportive care in British Columbia and the Yukon. This represents a catchment population of approximately 4.6 million. The model in Chapter 5 is applied at a cancer hospital located in Vancouver, British Columbia.

The manner in which the research of each chapter has been applied at the partner hospital varies. In Chapters 6 and 7 the best policies that were found in the course of the research were implemented at the partner hospitals and are used on a daily basis. The models of Chapters 4 and 6 were implemented in decision support software and are being used by hospital managers and staff. In particular, the model of Chapter 6 is implemented in third-party commercial software (see

Section 6.4). The results from Chapters 3, 4 and 5 led to management recommendations and practice guidelines for achieving certain objectives.

Although the models are motivated by cancer care, which is a specialized service, all models, except that of Chapter 7, are directly applicable in general hospital settings as well. This is particularly apparent in the involvement of Information Builders, a developer and distributer of business intelligence software. Information Builders developed the model of Chapter 6 into commercial software which is designed for use in specialty hospitals (i.e. cancer hospitals) and in non-specialty hospitals.

1.5 Summary of content

In Chapter 2 we review quantitative health care literature to illustrate the extent to which models encompass multiple hospital departments. We provide a general overview of the relationships that exist between major hospital departments and describe how these relationships are accounted for by researchers.

Chapter 2 is based on the following articles:

- Vanberkel P.T., Boucherie R.J., Hans E.W., Hurink J.L., Litvak N. (2010) *A survey of health care models that encompass multiple departments.* International Journal of Health Management and Information, 1 (1), 37 - 69
- Vanberkel P.T., Hans E.W. (2009) *Holistic healthcare modeling. A viewpoint on managing the complete patient care chain.* Contribution to the book: Operational Research Applied to Health Services in Action (ISBN 978-83-7493-409-1)

In Chapter 3 we address the decision of choosing a patient mix that leads to the most beneficial treatment case mix. We illustrate how

capacity, case mix and patient mix decisions are interrelated and how understanding this complex relationship is crucial for achieving the maximum benefit from a fee-for-service financing system. Studies to determine the case mix with maximum benefit exist in literature, however the hospital actions necessary to realize this case mix has received less attention.

We model the hospital as an $M/G/\infty$ queueing system to evaluate the impact of accepting certain patient types. Using this queueing model to generate parameters, an optimization problem is formulated. We propose two methods for solving the optimization problem. The first is exact but requires an integer linear programming solver whereas the second is an approximation relying only on dynamic programming.

The model is applied to the department of surgery at NCI. The model determines which patient types result in the desired growth in the preferred surgical treatment areas. The case study highlights the impact of striving for a certain case mix without providing a sufficiently balanced supply of resources. In the case study we show how the desired case mix can be better achieved with certain capacity investments.

Chapter 3 is based on the following article:

- Vanberkel P.T., Boucherie R.J., Hans E.W., Hurink J.L. (2011) *Optimizing the strategic patient mix.* Memorandum 1935, Department of Applied Mathematics, University of Twente, Enschede. ISSN 1874-4850

In Chapter 4 we address the decision of whether (and to what extent) to pool resources within a hospital. Hospitals traditionally pool resources into centralized functional departments such as diagnostic departments, ambulatory care centres, and nursing wards. In recent years this organizational model has been challenged by the idea that higher quality of care and efficiency of service delivery can be achieved when services are organized and focused around patient groups. Examples include specialized clinics for breast cancer patients and clinical pathways for diabetes patients. This leads to the question of whether

to become more centralized to achieve economies of scale or more decentralized to achieve economies of focus. In this chapter we examine service and patient group characteristics to study the conditions where a centralized model is more efficient, and, conversely, where a decentralized model is more efficient.

This relationship is examined analytically with a slotted queuing model to determine the most influential factors and then with a simulation to fine-tune the results. The tradeoffs between economies of scale and economies of focus measured by these models are used to derive general management guidelines.

The model is applied in the chemotherapy day unit (CDU) at NCI. The study investigates the expected service performance associated with a proposal to reallocate resources from a centralized (pooled) chemotherapy department to a breast cancer focused factory (unpooled). We show that a decrease in performance is expected and calculate the amount of additional resources required to offset these losses.

Chapter 4 is based on the following articles:

- Vanberkel P.T., Boucherie R.J., Hans E.W., Hurink J.L., Litvak N. *Efficiency evaluation for pooling resources in health care.* OR Spectrum (forthcoming)
- Vanberkel P.T., Boucherie R.J., Hans E.W., Hurink J.L., Litvak N., van Lent W.A.M., van Harten W.H. (2009) *Reallocating resources to focused factories: A case study in chemotherapy.* International Perspectives on Operations Research and Health Care: Proceedings of the European Working Group on Operational Research Applied to Health Services

In Chapter 5 we address the panel size decision for a single oncologist's practice (i.e. an unpooled practice). Panel size is defined as the number of patients that a physician can effectively be accountable for. Typically this is studied in general practice settings where general practitioners want to know how 'big' their practice can be before

the waiting times for appointments becomes too long or overtime too frequent. Panel sizes in a hospital environment have been studied less frequently, although there are similar concerns. The characteristics of a hospital which distinguish it from a general practice include higher turnover rates of patients and multiple patient and appointment types. We extend the earlier panel size models to account for these differences by modelling the panel size as the sum of patient type specific random variables.

We formulate two queueing network models to define the panel size random variables. The first queueing model is a multi-class open queueing model that assumes a stationary setting. This model represents an established oncologist with a mature clinic, who sees on average the same number of new patients per month. The second queueing model is a network of $M/G/\infty$ queues with a non-stationary arrival rate. The non-stationary setting is used to represent a new oncologist's practice.

We apply the panel sizing model in an oncology clinic that is part of the BCCA. We determine for a given patient mix the number of patients that can be seen in a stationary setting (i.e. by an established oncologist) and in a non-stationary setting (i.e. by a new oncologist). We combine both models to support long term capacity planning decisions.

Chapter 5 is based on the following article:

- Vanberkel P.T., Puterman M.L., Litvak N. *Panel sizing in oncology.* Working paper

In Chapter 6 we address the decision of how to schedule surgical specialties such that inpatient wards are not overwhelmed with patients. No other department influences the workload of a hospital more than the Department of Surgery, and in particular the activities in the operating room. These activities are governed by the master surgical schedule (MSS), which states which patient types receive surgery on which day. In this chapter we describe an analytical approach to project the

workload for downstream departments based on this MSS. Specifically the ward occupancy distributions, patient admission/discharge distributions, and the distributions for ongoing interventions/treatments are computed. Recovering after surgery requires the support of multiple departments such as nursing, physiotherapy, rehabilitation, and long-term care. With our model, managers from these departments can determine their workload by aggregating tasks associated with recovering surgical patients.

We model a patient's day-to-day recovery with binomial distributions reflecting the day of recovery and the responsible surgical specialty. Given that patients do not interfere with each other during their recovery, we model cohorts of patients independently. We then add the resulting bed usage of each cohort (with discrete convolutions) to compute ward occupancy distributions and other workload metrics.

The model was used to support the development of a new MSS at NCI and provides the foundation for a decision support system. After evaluating and discussing a number of proposals, a new MSS was chosen which was acceptable to operating room staff and which balanced the ward occupancy. After implementing the new MSS, a review of the bed use statistics validated the results.

The model has also been implemented in WebFOCUS, a commercial business intelligence software developed by Information Builders. The software supports health care managers in developing MSSs by reporting, for example, patient waiting times and ward occupancies. The method used by the software to predict the ward occupancies associated with each MSS proposal is based on the research of this chapter.

Chapter 6 is based on the following articles:

- Vanberkel P.T., Boucherie R.J., Hans E.W., Hurink J.L., van Lent W.A.M., van Harten W.H. *An exact approach for relating recovering surgical patient workload to the master surgical schedule.* Journal of the Operational Research Society (forthcoming)
- Vanberkel P.T., Boucherie R.J., Hans E.W., Hurink J.L., van Lent

W.A.M., van Harten W.H. *Accounting for inpatient wards when developing Master Surgical Schedules.* Anesthesia & Analgesia (forthcoming)

In Chapter 7 we address the decision of which (and to what extent) expensive medications should be made in advance of patient appointments. We investigate the impact that pharmacy medicine preparation policies have on patient waiting times. The chapter evaluates whether a reduction in waiting time resulting from medication orders being prepared in advance was justified, given that medications prepared in advance have a risk of being wasted if patients arrive too sick to receive treatment.

We derive explicit expressions to approximate patient waiting times and wastage costs, allowing management to understand the tradeoff between these two metrics. The explicit expressions allow the analysis to be easily repeated when medication costs change or when new medications/protocols are introduced. Using a case study and a simulation model, the approximations are evaluated.

This model was applied in the CDU at NCI and resulted in a new policy at the cancer centre which is expected to decrease the waiting time by half while only increasing pharmacy's costs by 1-2%.

Chapter 7 is based on the following article:

- Masselink I.H.J., van der Mijden T.L.C., Litvak N., Vanberkel P.T. (2010) *Preparation of chemotherapy drugs: planning policy for reduced waiting times.* Memorandum 1925, Department of Applied Mathematics, University of Twente, Enschede. ISSN 1874-4850

In Chapter 8 we offer concluding remarks and discuss future directions for operations research in health care.

Chapter 2

Survey of literature

Contents

2.1 Introduction

In the 1980s it became clear that the reductionist method made famous by F.W. Taylor was causing the American manufacturing industry to lose perspective of their overall factory [95]. The approach, which focused principally on analyzing individual components, failed to accurately account for their interactions. This narrow view was further compounded by the academic community which thrived on using reductionism for analyzing complex systems, ever the while increasing the gap between their research and actual practice [95].

Similar findings have been expressed about health care operations. The following excerpt from [36] provides a summary with examples.

> "In my experience, one of the major causes of inefficiency in the health care system is what I call 'localized expertise.' People working in the health care system are very knowledgeable about their own area but have relatively little understanding of what goes on in the next department. Doctors and nurses in the emergency department or in operating rooms do not really understand or sympathize with the problems faced by ward staff. People in hospitals have little appreciation for issues in long-term and home care. Occasionally, there are issues about 'my work is more important than yours' or 'my problems are bigger than yours.' More often, it is simply too difficult for people to get a real handle on the whole 'system.' This is where Operational Research professionals can play an important role".

In this review we want to deal with the potential of operations research in more detail; the scope of health care models is examined to determine the extent in which modellers account for the complex interdepartmental relationships that are inherent in health care. The chapter helps address the question: are researchers modelling hospitals in a way that reflects the reductionist view of managers or are they approaching hospital problems from a systems perspective?

All of the articles mentioned in this review are categorized in the online literature database ORchestra. ORchestra provides a comprehensive overview of scientific literature in the field of "Operations Research in Health Care" and can be accessed at [43]. ORchestra is maintained by the Center for Healthcare Operations Improvement and Research (CHOIR) at the University of Twente.

The chapter is organized as follows. Section 2.2 gives our definition for "holistic models" and describes the methodology used to identify relevant papers. Section 2.3 reviews models and broadly classifies them according to application area within the hospital. The chapter concludes with a discussion and summary in Section 2.4.

2.2 Defining holistic models

Jun et. al. [110] surveyed discrete-event simulation models in health care, citing over 100 articles and discussing the various applications in clinical settings. This widely cited paper "focuses on articles that analyse single or multi-facility health care clinics (for example, outpatient clinics, emergency departments, surgical centres, orthopedic department, and pharmacies)." With respect to patient flow and throughput, the paper identifies three areas of impact; how patients are admitted or scheduled; how patients are routed within the clinic; and how staff and resources are scheduled to match demand. They (Jun et. al. [110]) conclude that "despite the upward trend of health care simulation studies ... there is still a void in literature focusing on complex integrated systems" and suggest that this "may be due to the associated complexity issues and resource requirements."

From the results presented in [110], it is clear that prior to 1999 simulation was not widely used as a tool for modelling "complex integrated" health care systems. When one considers the advances in computers and simulation software since then, coupled with the ever increasing pressure on hospitals, it begs the question of whether this void has since been filled. In this section we investigate this question and focus on patient flow models with a scope that includes more than one department or unit. Although "more than one department or unit" is hardly a rigorous definition of a holistic health care model, it is thought that the vague, but inclusive, definition allows for a more complete review. In the interest of clarity, a short list of model types that are excluded from this review follows.

- Developing a surgical schedule and only considering resources within the department of surgery
- Models for medical decision making (e.g. comparing the effectiveness of different treatment approaches)
- Scheduling of physicians or hospital staff within a single department

As a starting point in identifying relevant literature, a list of all ar-

ticles citing [110] was compiled. Using Google Scholar, 70 articles were identified and classified as follows: 20 (28.6%) describe models containing more than one department or unit; 15 (21.4%) are instructional/tutorial in nature; seven (10.0%) are surveys; and 28 (40.0%) are applications and case studies within a single department or unit. The remaining papers mentioned in this chapter either cite or are cited by one of the 20 papers identified above as describing a model containing more than one department or unit. In total, the systematic review resulted in 88 articles describing models that encompass multiple hospital departments.

2.3 Common model scopes

Our review found that a substantial portion of the operations research health care literature focuses primarily on surgery, emergency medical care, inpatient wards, outpatient clinics and diagnostic imaging (DI), laboratory medicine (LAB) and pharmacy services. As such, the following subsections describe models found in these focal areas and the last subsection is used to describe other models which do not readily fit this broad classification. The importance and influence of each area on the hospital as a whole is discussed in the following paragraph.

The emergency department (ED), with its constant admission pressures [33], is often described as in a crisis [87] and has even been described as a threat to the future of the United Kingdom's National Health Service [24]. The surgery department and in particular "the master surgical schedule (MSS) can be seen as the engine that drives the hospital" [15]. The operation of both services depends heavily on the available capacity of the downstream inpatient wards. Prompt and efficient service within an outpatient facility can improve patient satisfaction [53, 99] resulting in patients being more likely to follow medical treatment plans [198] and thus reducing the need for patients to have surgery or visit the ED. Furthermore, DI services must be designed to examine patients with many different illnesses and "are

utilized by almost every category of patient which enters the hospital system. Hence, efficient utilization of x-ray facilities is a necessary condition for overall hospital efficiency" [156].

2.3.1 Emergency medical care

When one thinks of the emergency medical care, the ED is usually the first thing that comes to mind. However, there are a multitude of external groups supporting the ED including the upstream paramedics, the downstream wards, and the parallel stream support services such as DI, LAB and pharmacy. For a more detailed account of these and the many other service interactions within emergency medical care see [21, 74, 86].

Most operational research studies of the ED relate to waiting time and consider the layout of the ED, the prioritization of patients, and congestion. The models have "generally assumed that the processes outside the ED have little direct impact on its overall operations" [37]. However, studies without an operational research focus, such as [57,67], identified factors causing ED overcrowding that are outside of direct control of the ED. These factors are mainly lack of beds for patients admitted to the hospital, delays in service provided by DI, LAB and ancillary services, difficulty in arranging follow-up care and difficulty in the transfer process. Of the reviewed papers, only 12 describe models that explicitly account for processes outside of the ED. The scope of these models and the techniques used are discussed below.

All but one of the 12 papers explicitly consider the ED and ward relationship in their models. In [8,180], the authors use discrete event simulation to investigate the influence of the inpatient ward on ED waiting times. Using a systems dynamics approach, this relationship is also considered in [123]. In [37], the authors describe the use of simulation to analyze the cause and relationship of overcrowding in multiple EDs. The model described in [41] uses data mining techniques to identify ED/ward bottlenecks. The model described in [4] uses

discrete event simulation for a surgical ED which includes a regular-care unit, a semi-intensive care unit, and an intensive care unit (ICU). In addition to the wards, the models in [23, 49, 132, 170] consider the relationship between the ED and DI or the LAB.

A model with a slightly larger scope is described in [54]. Their model, although limited to cardiac care, incorporates both a normal inpatient ward and an ICU. By studying this relationship with queuing theory the authors contend that "raising occupancy rates of hospital management is unrealistic and counterproductive" and relate refused admission to the unavailability of downstream beds.

A "whole-system review of emergency and on-demand health care," described in [30], considers emergency medical care well beyond the boundaries of the ED. The focus is on the complete emergency health care system and therefore considers departments feeding the ED, such as ambulance services and primary care; downstream departments including wards and social services are also included. The systems dynamics model connects the departure rates (outflows) of one department with the arrival rates (inflow) of other departments, resulting in a model that is sensitive to the fact that a small change to one part of the system can have considerable impact elsewhere. With this robust model the authors are able to recommend a variety of admission practice approaches to reduce the demand for inpatient beds.

Table 2.1 summarizes the extent to which the papers in this subsection explicitly model the surrounding processes. It is not surprising that most of these papers include the downstream ward(s). Many studies claim that the lack of down stream beds is the "primary reason hospitals go into diversion" [103]. However, these studies and others [86], insist that all inputs and outputs be considered when addressing patient flow issues. While studies on congestion [57, 67] state that many of the causes are outside of the ED, this review only identified 12 models that explicitly account for interactions between the ED and adjacent departments.

Paper	Departments	Approach
[41]	ED, Wards	Data mining
[49]	ED, DI	Petri nets
[54]	ED, ICU, Wards	Queueing theory
[30]	Referrals, Ambulances, ED, LAB/DI, ICU, Wards	Systems dynamics
[180]	ED, Wards	Simulation
[37]	ED, Wards	Simulation
[23]	ED, LAB/DI, Wards	Simulation
[170]	ED, LAB/DI, Wards, OR	Systems dynamics
[132]	ED, LAB/DI, Wards	Simulation
[123]	ED, Wards	Simulation
[8]	ED, Wards	Simulation
[4]	ED, OR, ICU Wards	Simulation

Table 2.1 – The extent to which departments surrounding the ED are explicitly modelled

2.3.2 Surgical care services

Surgical care, like emergency care, does not operate in isolation, it "encompasses a continuum of activities through diagnostics, pre-operative, operative, and post-operative stages" [176]. Details on these activities are given in [160, 176]. For an up-to-date bibliography of operating room (OR) management articles see [58].

In the literature on surgical care service two themes are recurrent. First, a gate keeping system -the surgical schedule- is commonly used for adjusting the function of the OR. By changing when and what patients arrive, managers can control and possibly balance resource usage. For an overview of how hospitals develop this schedule see [20, 96, 190, 191, 197]. The second common theme is waiting list management. These models often consider how waiting patients are impacted by resources levels, resource distribution and patient priority schemes.

Scheduling systems, "which control the flow of patients into the sur-

gical arena, are frequently cited as a primary means of improving resource utilization" [130]. The development of a surgery schedule and the planning of patients is often described as a multistage process [14, 22] and, as is the case with ED models, often consider the impact of downstream bed availability. In [112], the authors examine the conflict created by elective patients being scheduled solely according to surgeon and OR availability under the assumption that an ICU bed will be available. The authors use a computer simulation to test a quota mechanism that aims to more evenly distribute the elective cases requiring admission to the ICU. In [37], the authors use simulation to model the patient flow starting from the surgical schedule and continuing through the OR, the recovery room, the ICUs and the regular inpatient wards. For various allocations of OR time, the model in [37] forecasts resulting beds and nursing levels. Using a mixed integer programming model described in [171], the authors show that "by reallocating the surgical specialties in the block schedule it is possible to reduce resource requirements needed to care for patients after surgery, while maintaining the throughput of patients."

A computer simulation that supports the care of patients with hip fractures is described in [50]. The simulation includes patient's presenting with a hip fracture, preoperative care, surgery, postoperative care, rehabilitation and discharge. The main objective of the model is to "simulate hip fracture care delivery reconfigured to comply with the national guideline on hip fracture care. This allowed exploration of how service changes affected outcomes and patterns of resource use." Using a multi-agent model it is possible for the service to explore "scenarios depicting varying degrees of guideline compliance."

Searching for articles that cite (or are cited by) the articles mentioned above revealed an extensive literature on the subject of OR scheduling. "A substantial and mature operations research literature describes techniques for manipulating the MSS, or the order of cases on the daily operating list, to maximize institutional goals or objectives" [22]. For an extensive bibliography on OR scheduling and planning see [34]. From this review of surgical scheduling models, it appears that studies

often consider a multitude of factors that are internal to the services, such as staffing and equipment, but usually only consider a single external factor, inpatient beds.

Other authors describe more general approaches to ensure the impact of the surgery schedule on adjacent processes is accounted for. In [176] a statecharts paradigm as a method "for constructing a discrete-event simulation model of the perioperative process" is presented. They argue this approach is a powerful method for "identifying likely responses to changes in the peri-operative process." In [15], the authors describe software for visually displaying the impact of the MSS on a compilation of dependent resources, including beds, human resources (e.g. nurses, anesthetists), specialized instruments and the radiology department.

Higher resource utilization and less surgery cancellations can result from careful scheduling and planning in the OR. This clearly has an impact on throughput and correspondingly on elective patient waiting times [34, 189]. However, waiting list management models for elective surgery often take the surgical schedule for granted and considers the allocation of resources (total OR hours and inpatient beds), and patient priority schemes as the variables [189]. These models are often specific to a surgical specialty [209], and are primarily used to quantify waiting list concerns, highlight imbalances in resources, or suggest ways to increase throughput. Outputs from the model may be used to support requests for a greater allocation of resources [19] or as decision support for selecting patients [71]. Waiting list management is further complicated by the social and political environment and the ethical implications [161] of using queues as a rationing device [84,134]. For a summary of waiting list practices and issues from a Canadian perspective see [19] and from a British perspective see [66, 208]. For a discussion on the appropriateness of patient priority schemes see [159].

In Table 2.2 a list of the identified papers relating the operation of the OR to surrounding departments is given. As was also the case in the preceding subsection, many authors explicitly model the downstream ward processes but represent upstream processes by statistical

distributions. The models discussed in this subsection all consider interactions with departments outside of the surgery department (the principle department under study). All recognize the importance of considering the availability of downstream ward capacity when making decisions in the OR.

2.3.3 Inpatient bed wards

The strong relationship between surgical care and emergency care and the inpatient wards is apparent from Tables 2.1 and 2.2 and the preceding subsections. It is not surprising, given this emphasis on wards and the fact that they are described as a hospital's "most expensive resources" [18], to find a portion of literature describing models with a focus solely on inpatient ward operations. What distinguishes the inpatient ward models from the models presented earlier is that these papers focus primarily on the inpatient bed resources.

A comprehensive simulation for bed capacity planning is presented in [90], which exposes the problems with hospital wide bed occupancy goals. "An acceptable occupancy, with its corresponding refusal rate, is a complex function of the patient case mix, the size of the bed compliment and the variability in patient [length of stay]". Similar sentiments are expressed in [54]. In [90], the authors also list 15 papers that address bed requirements using queueing models, integer programming, forecasting, or simulation and demonstrate the disadvantages of commonly used deterministic approaches. Other bed capacity studies consider critical care wards [26, 48, 155, 196], general inpatient wards [116], the distribution of beds [2, 154], the possibility of intermediate care wards [187] and controlling ward occupancy through admission practices [1, 195]. For further literature on bed capacity planning see [116].

Results presented in [44] state that queueing studies in health care are also often unit specific. "Although there is a vast literature available on the application of queuing theory in health care, none of the reviewed papers reported using queuing theory network models for systems of

Paper	Departments	Approach
[160]	OR, PACU, ICU	Mathematical programming
[136]	OR, Anesthesia	Statistical methods
[97]	OR, ICU	Mathematical programming
[158]	OR, ICU, Wards	Mathematical programming
[176]	Waiting lists, OR	Simulation
[183]	Waiting lists, OR	Mathematical programming
[140]	OR, PACU, Wards	Process reeningeering
[172]	Waiting lists, OR, Wards	Mathematical programming
[184]	Waiting lists, OR, Wards	Mathematical programming
[189]	Waiting lists, OR, Wards	Simulation
[106]	OR, PACU, Wards	Mathematical programming
[177]	OR, PACU	Statistical methods
[15]	OR, DI	Software
[29]	OR, Wards	Simulation
[32]	OR, PACU, Wards	Mathematical programming
[171]	Waiting lists, OR, ICU, Wards	Mathematical programming
[37]	OR, PACU, ICU, Wards	Simulation
[60]	OR, PACU, Wards	Statistical methods
[50]	OR, PACU, Wards, Rehab	Simulation
[85]	OR, Wards	Mathematical programming
[142]	OR, ICU	Statistical methods
[28]	OR, Wards	Simulation
[112]	OR, ICU, Wards	Simulation
[22]	OR, Wards	Mathematical programming
[59]	OR, ICU, Wards	Mathematical programming
[129]	Waiting lists, OR	Mathematical programming
[71]	Waiting lists, OR, Wards	Simulation
[162]	OR, DI	Simulation
[61]	Waiting lists, Clinic, OR	Software
[69]	OR, Materials management	Simulation
[114]	ED, ICU, OR	Simulation
[167]	OR, PACU	Software
[174]	OR, Wards	Mathematical programming
[209]	Waiting lists, OR, Wards	Simulation
[122]	OR, Wards	Simulation

Table 2.2 – The extent to which departments surrounding the OR are explicitly modelled

more than one unit" [44]. In two papers [44, 45], the authors use a step-by-step methodology "for analyzing hospital flow using queuing network and simulation models with the emphasis on solutions to peak flow periods." With the queuing theory model, the authors are able to find the system bottleneck and recommend resource levels for balancing utilization across the hospital. "Although [Queueing Network Analysis] was very effective in balancing the system quickly and easily, it has limitations. It does not consider time-dependence. It uses only the mean value of the length of stay in a unit bed ... it does not easily account for bed blocking." To combat this, a discrete event simulation is presented to provide insight into waiting times, throughput, and congestion. The advantage of hybrid queuing/simulation models is discussed in detail in [62, 64].

As a starting point, the hybrid queuing/simulation methodology described in [45] is applied to an obstetrics hospital because "it contains all of the features of a full service hospital but on a simpler scale." With the simulation model the authors are able to recommend how to "minimize blocking of beds from upstream units." A second study by the same authors [44] is of a 411 bed, 13 unit hospital, where patients are admitted via the ED, OR or direct admission to medical units (outpatient clinics are not included). The queuing analysis provided insight into bed balancing across the wards, while the simulation is used to maximize flow through the system.

Besides bed capacity decisions, the operation of inpatient wards is also studied. Typical impediments to patient flow in the inpatient wards, are outlined in [86]. In summary they include long patient discharge processes, long turn around times between patients, poor tracking of bed inventory and lack of information on new patients forcing wards to be reactive instead of proactive. Other anecdotal accounts of inefficiency made by ward staff to an author include: overworked staff underreporting available beds as a means to control workload, physicians keeping patients longer than necessary as a way of reserving beds and the inability of family members to pick up to-be-discharged patients in a prompt manner. Most models represent resources by

beds [82], demand by patient lengths of stay [194] and leave many of these operational issues unaddressed.

Even the best discharge planning does not help when there is no downstream capacity. Patients whose medical treatments are complete but cannot leave the hospital are often referred to as "alternative level of care patients" [13] or as "bed blockers" [168]. The cause of bed blocking can be "the reductions in numbers of beds in nursing homes, problems in funding from social service budgets, and waits for assessments from therapists or social services, for community services, or for equipment to be ordered, delivered, and installed" [18]. This problem is further compounded by poor coordination between the hospital and long term/social care, as discussed in [108]. The effect of bed blockers is often measured by the average fraction of beds occupied by patients whose medical treatment is complete. The range of this fraction has been reported as low as 0.5% [164] to as high as 35% [67,108]. Not surprisingly, other authors [67] found the effect of blocked beds was not limited to the wards and that the impact was also felt in the ED and critical care units where patients wait for admission to a bed. For a discussion on an initiative to integrate the hospital care with the nursing home care for elderly persons, see [13]. For a study relating bed blocking with community care and with the ED see [138]. Although this is clearly an area of importance for efficient use of inpatient beds it is not widely included in models of inpatient wards.

After examining three major portions of hospitals (emergency medical care, surgical care and the inpatient wards), we see an emphasis on the interaction between the wards and the ED and the OR. Five articles [113, 118, 123, 181, 209] consider the competing nature of the ED and the OR. This interaction, although perhaps not intuitive, is important because both services forward their inpatients to the wards. Even though many hospitals segregate their wards based on these services, it is often the case that they share beds at times of high demand, which happens to be the time of interest in most models.

The articles highlighted in this subsection are summarized in Table 2.3.

Paper	Departments	Approach
[118]	ED, OR, ICU	Simulation
[138]	Wards, Community care, ED	Data analysis
[44]	OR, ICU, Wards	Queueing theory & simulation
[45]	OR, ICU, Wards	Queueing theory & simulation
[13]	ED, Wards, Home care	Randomized controlled trial
[154]	Multiple Wards	Statistical methods
[196]	OR, ICU	Mathematical programming
[2]	Waiting lists, ICU, Wards	Queueing theory & simulation
[48]	ICU, Wards	Simulation
[155]	ICU, Wards	Statistical methods
[187]	Intermediate care, Wards	Queueing theory
[90]	ED, Wards	Simulation
[67]	ED, Wards	Simulation
[1]	OR, ICU, Wards	Mathematical programming
[123]	ED, OR, Wards	Systems dynamics
[113]	ED, OR, Wards	Simulation
[181]	ED, OR, Wards	Systems dynamics
[195]	OR, Wards	Process reengineering
[26]	ICU, Wards	Data analysis
[209]	ED, OR, Wards	Simulation

Table 2.3 – The extent to which departments surrounding the wards are explicitly modelled

2.3.4 Ambulatory care

The extent to which ambulatory care clinics are considered as part of a larger system is described in detail in [137]. The authors conclude that "despite the interrelatedness and the fact that patients are shared between facilities, outpatient care systems are rarely evaluated as a coordinated subsystem of a hospital." A rich literature on outpatient scheduling, albeit mainly focusing on a single department, started with [9] and is summed up in a comprehensive survey in [39].

Of the papers that cite [110], four are relevant to this subsection, since these models consider more than one department. All of these papers

describe models of ambulatory care centres, which are essentially clusters of outpatient services situated together. In [137] a comprehensive framework to measure the performance of "multi-facility outpatient centres" is presented. The paper includes a case study of an oncology centre which includes one surgical clinic, two medical clinics, one treatment clinic and 14 diagnostic testing facilities.

In [107], the authors describe a care centre which has multiple outpatient clinics located together and also managed as a single department. Their model uses a multi-class open queueing network and a simulation to model patient routing between the evaluation, x-ray, LAB, treatment and medication components in an urgent care centre. Their effort to achieve higher throughput by converting these processes from serial to parallel activities proved fruitless as the bottleneck activity (evaluation and re-evaluation by physicians) was the dominant cause of patient delays. The model described in [143] is for a muscul-skeletal unit, which the authors describe as "an innovative concept that was designed to integrate the activities of orthopaedics and rheumatology with specialist physiotherapy and podiatry." The authors describe five simulation models with the first four being typical what-if case studies applied within one hospital. In the fifth simulation the model is expanded to incorporate the "full integration of outpatient services across two hospitals" and is used to evaluate a new two-stage triage process.

The model described in [6] is different than the others in this subsection, since it describes a facility housing many ambulatory clinics, each of which has its own staff and appointment systems. The community based ambulatory care centre consists of seven services (ECG, dentistry, homeopath, chiropody, eye care, dietitian and family planning) in addition to four shared treatment rooms. Their simulation balances the patient loads of the groups and stimulates staff to "understand interactions across the whole picture, rather than just in the part that they would normally be involved with."

A final consideration for this subsection is the interaction of patients within the same department but at different stages of their care. As

Paper	Departments	Approach
[107]	Outpatient clinics, DI, LAB	Queueing theory & simulation
[137]	Outpatient clinics, DI, LAB, Pharmacy	Process reengineering
[143]	Multiple outpatient clinics	Simulation
[6]	Multiple outpatient clinics	Simulation

Table 2.4 – The extent to which departments surrounding the outpatient clinics are explicitly modelled

an example, most departments have the patient categories "new" and "return" for which the characteristic of the appointment may be different. This situation can be considered analogous to that of a patient visiting two different departments in which the outcome of the first appointment affects the second. Such a situation is investigated and discussed in [40]. The authors conclude "that patient sequencing has a greater effect on ambulatory care performance than the choice of an appointment rule, and that panel characteristics such as walk-ins, no-shows, punctuality and overall session volume, influence the effectiveness of appointment systems."

A summary for this subsection is given in Table 2.4.

2.3.5 Supporting departments

In this subsection three essential departments providing a supporting role in patient care are considered. Specifically, this subsection considers models for DI, LAB and pharmacy. For clarity we offer definitions of each department. The DI department interprets medical images such as x-rays, CT scans, nuclear medicine scans, mammograms and sonograms [47]. A typical LAB department consists of core-lab, microbiology, chemistry, blood transfusion services, and other hematological services. The pharmacy overseas the distribution of medication and ensures patients receive appropriate amounts of drugs and ensures that they do not interact. The involvement of pharmacy extends beyond

the walls of the pharmacy and includes consulting with staff during a inpatient's admission, stay, transfer and discharge [86]. For details on the operation of UK pharmacy systems see [56]. Other supporting services such as social work, physiotherapy and occupational therapy, are not considered in this review.

Of the over 70 papers citing [110] none describes a multi-departmental model with a focus on diagnostic services. This deficiency in literature is also noted in [31,74,156]. Without a single article as a starting point, the previously described methodology for searching literature had to be abandoned. In this subsection literature is identified by reviewing all articles that cited any of the six papers relating to radiology, hematology (LAB) and pharmacy discussed in [110].

Using a simulation model, [55] investigate the "relationship between the ward pharmacist's visit schedule and the delay between prescription of non-stock drugs and their delivery to the ward." The authors are cognizant of the fact that the distribution system is itself multidisciplinary and when changed, it affects "nursing and medical staff throughout the hospital as well as patients." For their case study the authors recommend the best time for pharmacists to visit the ward, and give a general conclusion that this best time can vary from ward to ward. Also using simulation, in [207], the medication ordering, dispensing and administration process is modelled to determine the potential benefits of replacing the paper based process with an automated system. The model described in [42] simulates a variety of scenarios to improve the working relationship between the OR and DI.

The operation of diagnostic services can be described as analogous to the operation of ambulatory clinics, particularly in terms of patient scheduling [39]. One difference however is that a coordinated approach is perhaps even more important for the overall patient care trajectory. Decisions on a patient's treatment may be placed on hold while waiting for the results from an x-ray, blood test or other test.

Table 2.5 summarizes the scope of the models discussed in this subsection.

Paper	Departments	Approach
[207]	Wards, Pharmacy	Simulation
[42]	OR, DI	Simulation
[55]	Wards, Pharmacy	Simulation

Table 2.5 – The extent to which departments surrounding DI and pharmacy are explicitly modelled

2.3.6 Geriatric care and mental health care

Three papers have been identified describing models which do not readily fit the classification scheme used in this chapter. Since they describe models that look at the system of care and not simply a single department in the care chain, we include them in our review. This subsection discusses these models, of which two are for mental health care and one is for geriatric care.

A model incorporating the various living situations of the mentally disabled in the Netherlands was developed in [119]. The "linear recursive stock flow model" is "developed from a dynamical systems point of view and incorporates the number of clients on the waiting list and the capacities of institutional and semi-institutional care." This macro level approach allows the entire system of residential care to be studied from a national perspective. Although the model is hampered by poor data, it did help pinpoint "critical elements in the waiting list discussion" and stimulated systems thinking by highlighting the effect of an increasing inflow and a stagnating outflow on patient waiting list.

A queuing theory model with blocking was developed in [115] to analyze the congestion in a mental health system. The model encompasses the interaction of the community, acute hospitals, extended acute hospitals, residential facilities and support housing. The analysis identifies the bottleneck resource and concludes that when planning, the transient behavior of the system is more importance than the steady-state. In their case study the authors find that "the shortage of a particular

Paper	Departments	Approach
[120]	Wards, Rehabilitation, Recuperative care	Simulation
[115]	Multiple mental health care facilities	Queueing theory
[119]	Waiting lists, Multiple mental health care facilities	Systems dynamics

Table 2.6 – The extent to which departments surrounding geriatric care and mental health care are explicitly modelled

type of facility may have created 'upstream blocking'. Thus removal of such facility-specific bottlenecks may be the most efficient way to reduce congestion in the system as a whole."

A "simulation study of a complex integrated health care system for older people, called intermediate care" is presented in [120]. When describing the scope to be studied the stakeholder "made it quite clear that they were keen to evaluate the whole intermediate care system and not just individual services." The system consisted of ten services, a community access rehabilitation team, a day hospital, a recuperative care service, and seven rehabilitation wards. Due to the complexity and the short time since the inception of intermediate care it was not exactly clear how these services interacted and/or complemented each other. In their paper the authors provide an extensive description of a soft systems methodology to first develop an understanding of the problem and then to determine a conceptual model. From this conceptual model a simulation was developed of the ideal system and then used to evaluate the utilization and to identify service gaps. Table 2.6 summarizes the scope of the models discussed in this subsection.

2.4 Discussion

Health care modelling literature is rife with studies on scheduling, resource utilization, and patient flow. However, these studies are often confined to the operation of a single department, ignoring many of the

complex relationships that exist between them.

This disjointed approach fails to offer coordinated patient trajectories and essentially represents a hospital as a collection of processes receiving patients from, and feeding patients into, buffers. For inpatients these effects and costs are high and direct, making the reduction of length of stay of patients a priority in hospitals and a common goal of many studies. For outpatients the costs associated with waiting for access to a service are not direct and often hidden. In addition to administrative costs, quality of life costs for patients cannot be understated. Besides the obvious extended period of time in poor health, there is anxiety associated with waiting, the possibility of further health deterioration, the loss of confidence in the hospital or physician, and furthermore, the compounded effect of all of these factors together.

As was shown in this chapter, some progress in this area is evident in the health care literature. Many models consider the impact of their operations on the downstream inpatient wards. Typical examples include bed occupancy being dictated by the OR schedule, and ED congestion being caused by an inability to admit patients to an already overcrowded ward. There is literature concerning a hospital's inability to discharge patients into long-term care. Hospitals are developing ambulatory care centres that locate multiple specialties together so that a patient's ambulatory treatment can, at the least, happen in the same space, and at the best, be efficiently coordinated. Pharmacy services identify that the drug distribution network is multidisciplinary and has significant impact on the work of physicians and nurses in addition to patient care implications.

The review contains 88 papers describing patient flow models that considered resources from two or more hospital departments. This amount is consistent with findings of other authors [74, 110] who concluded that although there is an abundance of models for health care processes, few consider multiple units or departments.

All of the 88 models include the *interactions with downstream departments*. This highlights that congestion in one department is often

related to an inability to forward patients to a downstream department.

Of the 88 models, 30 explicitly model the *interaction with upstream departments* (i.e. those departments from which their patients are received from). The remaining models use distributions to capture the variations associated with arrival patterns. Although this method is preferential to using only averages it fails to distinguish between the variation caused by the random nature of illnesses and the variation induced by preceding departments. Such oversight may result in implementing complex policies to deal with variation instead of eliminating it at the source.

Finally, only 13 of the 88 models consider the *interaction with parallel stream support services* (e.g. LAB and DI) and how these departments impact the flow of patients through the hospital. Although many patients require blood work, x-rays or other exams in order to be properly treated or diagnosed, very few models include their interactions with the main department under study.

Chapter 3

Patient mix optimization

Contents

3.1 Introduction

In recent years hospital financing has changed from a budget oriented (lump sum) system to a fee-for-service system in many jurisdiction [76]. This transformation is intended to enhance accountability and to motivate hospitals to become more efficient. Diagnosis related groups (DRGs), a concept which makes health care services a commodity, is facilitating this change. A DRG describes the whole spectrum of activities involved in treating a certain disease or condition. The reimbursement to the hospital for each DRG treatment is fixed, meaning hospitals that provide the treatment for lower costs can realize greater

profits, hence DRGs motivate efficiency. Variants of DRGs were introduced to achieve the same hospital financing transformation in many countries [178].

As a consequence of fee-for-service financing, hospitals must consider their DRG case mix and evaluate which services should be expanded and which should be discontinued [111, 199]. This is evident in papers [125,163] that focus on determining which DRGs are of the most benefit to the hospital. How to realize this DRG case mix has so far not been addressed in literature. Furthermore, empirical research indicates that hospitals struggle to make choices that lead to desired DRG case mixes [51, 52].

To achieve a desired DRG case mix, hospitals must entice certain patients to the hospital. Patients are usually referred to the hospital by a general practitioner (GP) who evaluates a patient's symptoms and decides whether the patient should see a specialist. A referral from a GP does not specify which DRG treatment is required but rather the symptoms and the most appropriate modality. The patient then meets a specialist who decides on a treatment plan and based on this treatment plan the corresponding DRGs are recorded. These DRGs may or may not be the ones of greatest benefit, however it is atypical to turn patients away at this point.

Through advertising and promotion to GPs, hospitals can encourage patients with certain diseases or symptoms to come to their hospital for treatment. However, knowing which symptoms will lead to the desired DRG case mix is not immediately obvious. Arrivals of patients (characterized by their symptoms) follow a stochastic process, and the required treatment can not be predicted with certainty. Determining, on the basis of symptoms, which types of patients (patient mix) to entice to the hospital in order to achieve the desired DRG case mix is the focus of this chapter.

An as example, consider the treatment of colorectal cancer. A patient suspected of having colorectal cancer is referred to a hospital for further testing. The results from the testing could lead to surgery,

chemotherapy, radiotherapy or palliative care for malignant cases, further screening for benign cases, a referral to a different oncology modality or even no further treatment. Within each of the treatment scenarios, there are several treatment options (i.e. DRGs) of which some are more desirable than others. Patient types in this example can be defined in many ways, but common factors indicating the prevalence of colorectal cancer include: personal or family history of polyps, history of colorectal cancer and/or bowel disease, ethnic background, diet, weight, alcohol use or smoking. Patient types can be further defined by symptoms such as constipation, diarrhea, blood in stool or jaundice. Patient types have uncertain arrival rates and with some probability require specific treatments. Thus choosing the best patient types to achieve the hospital's desired DRG case mix is not immediately obvious.

Hospitals are also constrained by their capacity levels which presumably relate to their desired DRG case mix. When capacity is overwhelmed the number of patients in the system increases, resources become more highly utilized, but patient access times become worse. In this research, to account for quality degradation due to demand exceeding capacity, we limit the fraction of time demand is allowed to exceed capacity.

In this chapter, we choose which patient types lead to a DRG case mix of maximum benefit over time. The chosen patient types are then "added" to the patient mix. Once added, patient types cannot be removed in future periods, as allowing such an "on-again, off-again policy" would create undesired confusion about the offerings of the hospital. In this way, our problem has properties similar to the project sequencing problem (PSP). The PSP determines which capacity expansion projects to implement in order to fulfill a growing demand for capacity.

We model the hospital as an $M/G/\infty$ queueing system and formulate an integer linear program (ILP) to exactly solve our problem. Using results from PSP literature we also formulate an approximate solution.

Statement of contribution: We develop a mathematical model to determine the policy for accepting new patient types that best matches the desired DRG case mix. To our knowledge it is the first time that capacity, DRG case mix and patient mix decisions are accounted for in a single model to facilitate joint decision making over a long term planning horizon.

The chapter is organized as follows. Section 3.2 formally defines and specifies the optimization problem and the queueing model. Section 3.3 introduces the PSP and illustrates how it can be used to approximately solve our problem. In Section 3.4 a case study is solved and the approximation evaluated. Throughout the chapter the terms DRG and treatment are used interchangeably.

3.2 Model description

The problem addressed in this chapter is as follows. Given that a hospital desires a certain DRG case mix, which patient types should be accepted (and when) to achieve this case mix while ensuring capacity restrictions are accounted for. We assume the relative importance of the DRGs are known and the capacity of the hospital to provide treatments is known for a finite time into the future. After a patient type is accepted, the number of arrivals of that patient type is modelled as a stochastic process. Upon arrival, a patient of a given patient type receives treatments according to some given probability distribution. Our model treats time as continuous and considers a finite planning horizon.

The formal problem description and a formalization as a combinatorial optimization problem are presented in Subsection 3.2.1. The calculation of some of the parameters of this combinatorial optimization problem is done using a queueing model which is described in Subsection 3.2.2.

3.2.1 Combinatorial optimization problem

Consider a set of patient types $P = \{1, 2, \ldots, N\}$ and a set of possible treatments $T = \{1, 2, \ldots, M\}$. A patient of type $n \in P$ has a probability $p_{n,m}$ of requiring treatment $m \in T$. The duration of treatment m has cumulative distribution $B_m(\cdot)$ with mean $\mathbb{E}[B_m]$. Let the number of arrivals of patient type n in period $[0, t)$ be specified by a given random variable $\Lambda_n(t)$ and let $G_m(t)$ be a given model input which describes the volume of ongoing treatments m for which the hospital has capacity for at time t.

For modelling the problem, we introduce variables $S_m(t)$ and $D_m(t)$ where $S_m(t)$ is the distribution for the number of patients receiving treatment m at time t and $D_m(t)$ is the distribution for the number of completed treatments m at time t. Note that $S_m(t)$ and $D_m(t)$ result from the choice of patient types to be accepted. The desired DRG case mix is reflected by values w_m, which specify the relative importance (or value) of treatment m.

The problem now is to indicate for each patient type the first moment in time t_n that patient type n is accepted. Note that for all times after t_n patient type n must also be accepted. Then the goal is to determine t_n such that the weighted number of treatments (weighted according to w_m) is maximized while ensuring that the number of treatments does not exceed $G_m(t)$ for more than a fraction φ_m of time. In other words, a hospital with capacity $G_m(t)$ wishes to maximize the weighted number of treatments they perform, whereby it is acceptable to exceed their capacity for a certain fraction $(1 - \varphi_m)$ of the time.

The value $\varphi_m \in (0, 1)$ is an input parameter reflecting the hospital's risk aversion for operating over capacity. A high φ_m value means demand will exceed capacity frequently (causing, for example, backlogged demand) whereas a low φ_m value means demand will exceed capacity less frequently (causing, for example, under utilized resources).

Let $\gamma = (t_1, t_2, \ldots, t_N)$ be a vector of chosen times to accept patient types n and let the resulting reward be measured by the discounted

weighted sum of completed treatments for decision γ. Discounting future costs by $e^{-\alpha t}$ (where $\alpha \in (0,1)$ is the discount factor) to time 0 ensures that later costs are adequately taken into account. Finding the optimal γ leads to the following optimization problem,

$$\begin{aligned} &\text{maximize } \int_0^T C_t(\gamma) e^{-\alpha t} dt \\ &\text{subject to } \mathbb{P}(S_m(t) \geq G_m(t)) \leq \varphi_m \qquad \forall m, t \end{aligned} \tag{3.1}$$

where,

$$C_t(\gamma) \;=\; \sum_{m=1}^{M} \mathbb{E}[D_m(t)] w_m. \tag{3.2}$$

Reward function (3.2) rewards according to the number of treatments completed, and is motivated by the financing structure at the hospital under study. Other choices are possible and the choice can be determined by the underlying decision process. Obvious choices include:

1. $C_t(\gamma) = \sum_{m=1}^{M} \mathbb{P}(S_m(t) \geq G_m(t))$ that rewards according to the fraction of patients that exceed capacity
2. $C_t(\gamma) = \sum_{m=1}^{M} \max\{\mathbb{E}[S_m(t) - G_m(t)], 0\}$ that rewards according to the expected amount by which capacity is exceeded
3. $C_t(\gamma) = \sum_{m=1}^{M} \mathbb{E}[S_m(t)]$ which rewards by the expected number of patients receiving treatment m (assuming $S_m(t)$ is appropriately constrained in relation to $G_m(t)$)

Capacity in our model ($G_m(t)$) is specified as the volume of ongoing treatments m that the hospital can accommodate at time t. This implies that the decision of how to allocate available resource *time* to treatment types, has already been made (again this follows from the hospital under study). For example, consider an MRI machine that is available for 2000 minutes per week for treatments A and B. Assume treatments A and B require 10 and 20 minutes respectively. One way

to divide the 2000 MRI minutes is to allocate 1200 minutes to A and 800 minutes to B, leading to a volume of $G_A(t) = 1200/10 = 120$ A treatments and $G_B(t) = 800/20 = 40$ B treatments per week. Because there are numerous ways to distribute the 2000 minutes between the two treatment types, there are numerous possibilities for $G_A(t)$ and $G_B(t)$.

Whereas in our model, the decision of how much time to allocate to each treatment type is already made i.e. we have only a single value of $G_A(t)$ (and a single value of $G_B(t)$), in other settings, it may be desirable to have this resource allocation decision be part of the model. The presented model can be generalized to accomplish this by defining capacity by the available treatment time of a resource. In this case the demand for resources can be determined by multiplying the volume of ongoing treatments $S_m(t)$ by the resource time needed for treatment m.

To solve (3.1), we have to specify how random variables $S_m(t)$ and $D_m(t)$ can be determined for a given vector of γ. The queueing model defined in Subsection 3.2.2 is used for this purpose.

3.2.2 Queueing model

To calculate the number of patients in treatment ($S_m(t)$) and the number of patients completing treatment ($D_m(t)$), we model the hospital as an $M/G/\infty$ queueing system. Since the population is a large, and patients get ill independently of each other, it is natural to assume that they get ill according to a Poisson process $\tilde{\Lambda}_n(t)$. Furthermore, since trends and seasonality are common in many diseases it is also natural to assume a non-stationary Poisson process. It follows that patients of type n that arrived to the hospital have a Poisson distribution $\Lambda_n(t) = \tilde{\Lambda}_n(t)F_n(t)q_n$ with mean $\lambda_n(t)$ where q_n is the fraction of the population choosing the considered hospital and where $F_n(t) = 1$ when patient type n is accepted at time t (i.e. $t_n \geq t$) and $F_n(t) = 0$ otherwise (without loss of generality we assume $q_n = 1$).

With probably $p_{n,m}$ [or $p_{n,m}(t)$] a patient of type n requires treatment m and, thus, the arrival process of patient type (n, m) is Poisson at rate $\lambda_{n,m}(t) = \lambda_n(t)p_{n,m}$. It can readily be seen that the number of patients of type n requiring treatment m at time t, $S_{n,m,t_n}(t)$, is distributed as the number of customers in a non-stationary $M/G/\infty$ queue with arrival rate $\lambda_{n,m}(t)$ and cumulative service time distribution $B_m(\cdot)$ and is given by (see [135]),

$$\mathbb{P}(S_{n,m,t_n}(t) = s) = \frac{e^{-\mathbb{E}[S_{n,m,t_n}(t)]}(\mathbb{E}[S_{n,m,t_n}(t)])^s}{s!} \tag{3.3}$$

where,

$$\mathbb{E}[S_{n,m,t_n}(t)] = \mathbb{E}[\lambda_{n,m}(t - B_{e,m})]\mathbb{E}[B_m] \tag{3.4}$$

and $B_{e,m}$ is the excess service time with the following cumulative distribution function,

$$\mathbb{P}(B_{e,m} \leq t) = \frac{1}{\mathbb{E}[B_m]} \int_0^t (1 - B_m(u))\, du.$$

Observe that $S_m(t) = \sum_{n=1}^{N} S_{n,m,t_n}(t)$ is the sum of Poisson random variables and therefore is Poisson distributed with rate parameter $\mathbb{E}[S_m(t)] = \sum_{n=1}^{N} \mathbb{E}[S_{n,m,t_n}(t)]$. Like $S_m(t)$, $D_m(t)$ is also Poisson distributed in a non-stationary $M/G/\infty$ queue. In particular, the number of completed treatments m of patient type n at time t is Poisson with mean $\mathbb{E}[D_{n,m,t_n}(t)] = \mathbb{E}[\lambda_{n,m}(t - B_m)]$ and $\mathbb{E}[D_m(t)] = \sum_{n=1}^{N} \mathbb{E}[D_{n,m,t_n}(t)]$.

3.2.3 Model characteristics

In this subsection, some observations about infinite server queues and our optimization problem (3.1) are explained. These observation are exploited later when solving the problem. First, entities (patients in our case) are independent of each other in an infinite server queue. This means, for example, that the amount of time in the system is

not impacted by other patients or the order in which they arrive. It follows, that the total reward for accepting a new patient type can already be computed at its acceptance time t_n, since these patients are not influenced by other patients, and therefore not influenced by future decisions. It follows that the reward for adding patient type n, discounted to time t_n is,

$$c_{n,t_n} = \sum_{m=1}^{M} \left(\int_{t_n}^{T} \left(\mathbb{E}[D_{n,m,t_n}(t)] e^{-\alpha t} \right) dt \right) w_m$$

and it follows that the objective function from (3.1) can be rewritten as,

$$\int_0^T C_t(\gamma) e^{-\alpha t} dt = \sum_{n=1}^{N} c_{n,t_n} e^{-\alpha(t_n)}.$$

The second observation about (3.1) is that the constraints are expressions for the φ_m percentiles of distributions $S_m(t)$ which, when $S_m(t)$ is known, can be evaluated exactly. Let $\mathbb{Q}[x]$ represent the φ_m percentile of distribution x, then,

$$\begin{aligned} \mathbb{Q}[S_m(t)] &= \sup\{s : \mathbb{P}(S_m(t) \geq s) \leq \varphi_m\} \\ &= \sup\left\{ s : \sum_{k=1}^{s} \frac{\mathbb{E}[S_m(t)]^k e^{-\mathbb{E}[S_m(t)]}}{k!} \leq \varphi_m \right\}. \end{aligned}$$

Using this observation the constraints from (3.1) can be rewritten as,

$$\text{subject to } \mathbb{Q}[S_m(t)] = \mathbb{Q}\left[\sum_{n=1}^{N} S_{n,m,t_n}(t) \right] \leq G_m(t) \qquad \forall m, t.$$

3.2.4 Discrete time formulation

In this subsection we formulate a discrete time version of (3.1) which eliminates the need to evaluate the integral in its objective function.

In the discrete time version of our problem the continuous time discount equation $e^{-\alpha t}$ is replaced by the equivalent discrete time discount equation $(1/(1+\alpha))^t$. Let $x_{n,t}$ be binary decision variables reflecting the first moment in time when patient type n is accepted. For example, when $x_{3,5} = 1$ this means $t_3 = 5$. This leads to the following ILP problem,

$$\text{maximize} \sum_{n=1}^{N} \sum_{t=0}^{T} c_{n,t_n} x_{n,t} \left(\frac{1}{1+\alpha}\right)^t$$

subject to

$$\sum_{t=0}^{T} x_{n,t} \leq 1 \quad \forall n \tag{3.5}$$

$$\sum_{\tau=0}^{t} \mathbb{Q}\left[\sum_{n=1}^{N} S_{n,m,t-\tau}(t)\right] x_{n,t-\tau} \leq G_m(t) \ \forall m,t \tag{3.6}$$

$$x_{n,t} = 0 \text{ or } 1 \quad \forall n,t.$$

Constraints (3.5) ensure that each patient type is accepted only once, while constraints (3.6) ensure that the φ_m percentile of demand for treatment m does not exceed $G_m(t)$. For moderate sized instances the resulting ILP can be solved by commercial ILP solvers like e.g. ILOG CPLEX Solver.

3.3 Approximate solution approach

In this section we introduce an approximate solution approach (ASA) to our problem which does not rely on an ILP solver. The approximation exploits the structure of our problem which is similar to the well studied project sequencing problem (PSP). The PSP can be solved directly with dynamic programming [70]. As explained in this section, although our problem is similar to a PSP, it is too general to be solved directly with dynamic programming. To overcome this, we relax one

of the elements of our problem. The relaxation amounts to ignoring the time required for the newly accepted patient type to populate the system i.e. we ignore the "startup" time. Formally this is achieved by changing the definition of $S_m(t)$ (See Subsection 3.3.2). Furthermore we ignore one of the conditions of the dynamic program necessary to obtain an optimal solution. We argue that although the solution is not guaranteed to be optimal, within the range of typical problems, ignoring this assumption has little impact on the solution quality (as illustrated with numeric examples in Section 3.4).

The steps in the approximation are as follows. We first relax our problem (by defining $S_m(t)$ differently) so that it is a PSP. We then apply the dynamic program of [70] to determine the sequence in which the patient types should be accepted, i.e. the order not the time. Once this sequence is known we use the original definition of $S_m(t)$ to iteratively determine the best time to accept each patient type for the given sequence.

Using the ASA to solve (3.1) has three distinct advantages over the ILP approach. The first being that the ASA relies on dynamic programming and does not require a costly ILP solver. The second is that the ASA separates the sequencing decision from the timing decision and as such reduces the complexity. Hence large problem instances that are intractable for the ILP can be approximated using the ASA. The third advantage is that the PSP is a commonly studied problem and, as such there are heuristics and extensions to the dynamic programming approach to solve even larger problem instances, see [98,117]. By formally defining the relationship between our problem and a PSP, we are able to leverage existing PSP literature to approximately solve very large problem instances.

This section is organized as follows. In Subsection 3.3.1 we introduce the PSP and discuss the dynamic programming solution approach. In Subsection 3.3.2 we formally introduce the relaxation of our problem needed for this approach. In Subsection 3.3.3 we formally define the ASA to our problem.

3.3.1 Project sequencing problem review

There is a rich literature on capacity expansion problems, as indicated by a number of reviews [131,148] and books [77,94] spanning multiple decades. Broadly stated, the literature concerns choosing the timing, size and type of capacity expansion needed to fulfill growing demand. Typically, the objective is to minimize the discounted cost of meeting demand and in doing so, finds the optimal balance between expanding in large increments to achieve economies of scale, verses small increments which reduce opportunity (or excess capacity) costs. The literature addresses many variants of the problem, including finite / infinite planning horizons, linearly /non-linearly growing demand, deterministic / stochastic demand, and continuous / discrete expansion sizes. When the choice of expansion projects is limited to a finite set of projects, each with a specified expansion size, the problem is called a PSP.

The PSP problem [70, 131, 152], assumes a finite set of projects $R = \{1, 2, ..., Y\}$ where each project $y \in R$ has an implementation cost v_y and capacity z_y. The objective is to determine the implementation times (t_y) for each project which minimizes the discounted cost while fulfilling a pattern of deterministic demand $s(t)$. Formally the problem can be stated as follows,

$$\begin{aligned} &\text{minimize } \sum_{y \in R} v_y e^{-\alpha t_y} && (3.7)\\ &\text{subject to } \sum_{y \in R} z_y \delta_y(t) \geq s(t) \qquad \forall t \end{aligned}$$

where $\delta_y(t) = 1$ if $t \geq t_y$ (and $\delta_y(t) = 0$ otherwise).

To solve (3.7), Erlenkotter [70] proposes a backward dynamic programming model to determine the optimal implementation sequence. From this sequence, the optimal implementation time for each project can be derived. Let X be the state variable where X describes a subset of R representing the set of implemented projects. Let $\tau(X)$ be the

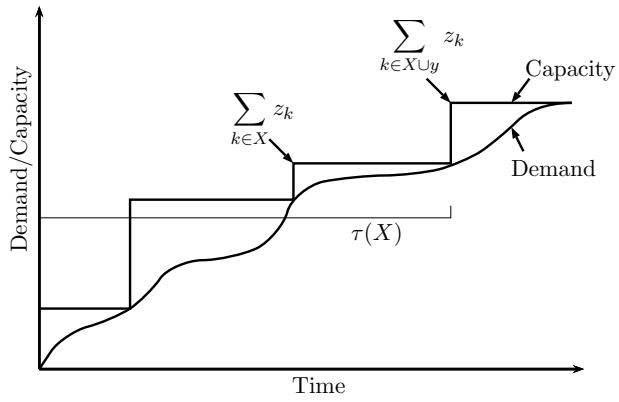


Figure 3.1 – The capacity expansion process

latest time when the capacity available from the implemented projects is greater than the demand, i.e. $\tau(X) = \max\{t | \sum_{y\in X} z_y \geq s(t)\}$.

When minimizing excess capacity costs (and when the discount rate α is positive), it is optimal to add additional capacity only when existing capacity is exhausted [70, 146, 147, 152]. Figure 3.1 displays the capacity expansion process for a one dimensional problem. The following backward dynamic program determines the optimal sequence for implementing the Y projects,

$$\begin{aligned} f(R) &= 0 \\ f(X) &= \min_{y\notin X} \left(v_y e^{-\alpha\tau(X)} + f(X \cup y) \right), \quad \forall X \subset R. \end{aligned} \tag{3.8}$$

A network representation of (3.8) can be constructed. Figure 3.2 displays an example with $Y = 3$ projects. Finding the shortest path from node $\emptyset$ to $\{1, 2, 3\}$ also gives the optimal sequence in which the projects should be implemented.

Once the sequence is known, the time to implement each project can be determined iteratively using $\tau(X)$. When demand is linear, the

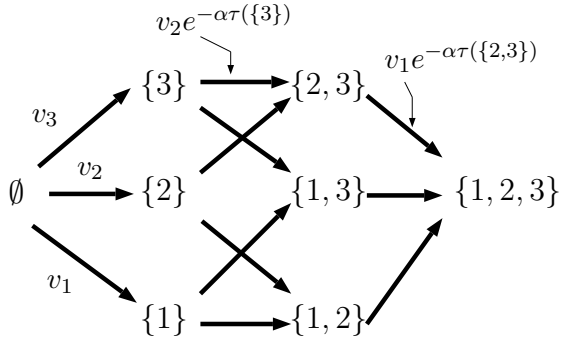


Figure 3.2 – A network representation of the PSP

optimal policy can be determined without dynamic programming as described in [70]. Other solution approaches involve integer programming techniques which assume a discrete time scale [151, 182] and "select jointly the choice of expansion and its timing" [70].

Under the following five assumptions, the Erlenkotter dynamic programming approach finds the optimal solution:

1. The full capacity is available for use instantaneously upon completion of a project
2. Project capacity once created has infinite life and does not change over time
3. a) The investment cost is incurred at the time the project is completed b) The investment cost does not vary with time
4. Variable operating and distribution costs are proportional to the amount actually produced and identical for all projects
5. Demand must be supplied from current production

3.3.2 Analogy between our problem and the PSP

By considering each patient type n as a project, our problem is similar to a PSP. Note however, that we are maximizing the reward, not minimizing capacity investments and that we restrict this reward by

available capacity $G_m(t)$ which is independent of the decision variables. In our problem $S_m(t)$ depends on the decision variables, thus, when comparing our problem to the PSP, $\mathbb{Q}[S_m(t)]$ is akin to "capacity" and $G_m(t)$ is akin to "demand".

As discussed below, we can not use Erlenkotter's dynamic program directly to solve our problem because it violates assumption 1. To overcome this we modify (relax) our problem. Assumption 2 is also violated, however as discussed below, the impacted of violating this assumption is minimal. The remaining 3 assumptions are satisfied in our problem.

1. The full capacity is available for use instantaneously upon completion of a project: This assumption is needed so that the total capacity of a given state can be determined from the state description. It follows that since the state indicates only the implemented projects (and not *when* each project was implemented) that the capacity of a project be independent on how much time has passed since it was implemented. In other words the capacity must be determined independently of the time of implementation. Using the definition of $S_{n,m,t_n}(t)$ from Subsection 3.2.2, this is not possible because $\mathbb{E}[S_{n,m,t_n}(t)]$ depends on t_n, see (3.4).

To overcome this, and to allow the use of dynamic programming we relax our definition of $S_{n,m,t_n}(t)$. In our relaxation, we model the number of patients in the system with a steady state M/G/∞ queue [83]. In such a queue, the number of patients in the system ($S'_{n,m}(t)$) is Poisson distributed with mean $\mathbb{E}[S'_{n,m}(t)] \approx \lambda_n(t)p_{n,m}\mathbb{E}[B_m]$. Notice that the definition of $\mathbb{E}[S'_{n,m}(t)]$ is independent of t_n. Formally the relaxation that we use in the ASA is as follows,

$$\mathbb{E}[S_{n,m,t_n}(t)] \approx \mathbb{E}[S'_{n,m}(t)] \approx \lambda_n(t)p_{n,m}\mathbb{E}[B_m]. \tag{3.9}$$

Obviously, in a similar manner $S'_m(t)$ denotes the volume of patient receiving treatment m at time t in this relaxed setting. $S'_m(t)$ is Poisson distributed with mean $\mathbb{E}[S'_m(t)] = \sum_{n=1}^{N} \mathbb{E}[S'_{n,m}(t)]$.

This modification is consistent with other applications of the PSP

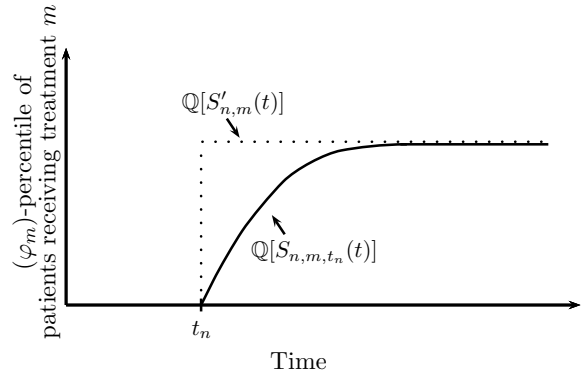


Figure 3.3 – Illustration of $\mathbb{Q}[S_{n,m,t_n}(t)]$ and $\mathbb{Q}[S'_{n,m}(t)]$

where a project's capacity needs to be brought online gradually. Furthermore, when the time between accepting new patient types is much longer than the mean service time, one would expect that $\mathbb{E}[S_{n,m,t_n}(t)]$ will have converged before the next patient type is accepted and thus this assumption does not impact the solution. In Figure 3.3 a sketch of the difference between $\mathbb{Q}[S_{n,m,t_n}(t)]$ and $\mathbb{Q}[S'_{n,m}(t)]$ is given.

2. Project capacity once created has infinite life and does not change over time: Since $\mathbb{Q}[S'_{n,m}(t)]$ depends on $\lambda_n(t)$ which is non-stationary in time, this assumption is violated. However, given that the prevalence of a disease (i.e. the arrival rate for a population of patients) changes gradually over time, one expects that the impact of violating this assumption is minimal.

3. a) The investment cost is incurred at the time the project is completed: Since the patients are assumed to be independent, the reward can be accounted for at the moment in time when the patient type is accepted. Although the rewards will take place in the future, they are known at time at, and can be discounted to, time t_n.

3. b) The investment cost does not vary with time: This is true in our case since the reward per treatment (w_m) does not change over time.

4. Variable operating and distribution cost are proportional to the amount

actually produced and identical for all projects: In our problem there are not any variable operating or distribution cost.

5. Demand must be supplied from current production: Our problem is equivalently constrained.

3.3.3 Dynamic programming formulation

In this subsection we first describe how the Erlenkotter approach is used to determine the optimal *sequence* to add patient types, given $S'_{n,m}(t)$. We then use the original definition of $S_{n,m,t_n}(t)$ to determine the *times* t_n and the *reward* associated with the sequence. These steps together constitute the ASA.

Sequence in which to add patient types: Let I be a subset of P indicating the patient types n that have already been accepted at a given moment in time. Our objective is essentially to maximize treatments, and since our discount rate is positive, it is optimal to accept additional patient types as soon as possible, i.e. at the first moment in time when a patient type can be added such that $\mathbb{Q}[S'_m(t)] \leq G_m(t)$ for all m. When the set of patient types I is already accepted and when patient type $r \notin I$ is to be added, then let this point in time be $\tau(I \cup \{r\})$ which is computed as follows,

$$\tau(I \cup \{r\}) = \inf\{t | \mathbb{Q}[S'_m(t)] \leq G_m(t),\ \forall m\}. \tag{3.10}$$

As explained in Subsection 3.2.1, the reward for accepting a patient type is the discounted weighted sum of the completed treatments. For the dynamic program to work, we must compute this reward at the moment in time when the patient type is first accepted. When the set of patient types I are already accepted and when patient type r is to be added, then the total reward (discounted to time $\tau(I \cup \{r\})$) for adding patient type r is,

$$c_r = \sum_{m=1}^{M} \left(\int_{\tau(I \cup \{r\})}^{T} \mathbb{E}[D_{r,m,\tau(I \cup \{r\})}(t)] e^{-\alpha t} dt \right) w_m. \tag{3.11}$$

Using the following backward dynamic program, the sequence in which to add patient types can be determined,

$$\begin{aligned} f(P) &= 0 \\ f(I) &= \max_{r \notin I} \left(c_r e^{-\alpha \tau(I \cup \{r\})} + f(I \cup \{r\}) \right) \quad \forall I \subset P. \end{aligned} \tag{3.12}$$

Let $\vartheta = \{x_1, x_2, \ldots\}$ be this sequence, where x_i is the i^{th} patient type accepted. For example when $\vartheta = \{6, 2, 4, \ldots\}$ then patient type 6 should be accepted first, followed by type 2 and type 4, etc.

Implementation times and reward: Once ϑ is known it is possible to determine the implementation times and the overall reward associated with this sequence. One option to determine the implementation times and the overall reward, is to add the following constraints to the ILP formulation of Subsection 3.2.4,

$$t_{x_k} \leq t_{x_{k+1}}, \quad k = 1, 2, \ldots, N.$$

Of course this means the ASA requires an ILP solver and thus, this is not a desirable option. Alternatively, it is possible for a given sequence to compute the implementation times iteratively as follows. Start initially with $t_{x_1} = t_{x_2} = t_{x_3} = \ldots = \infty$ and then use (3.13) to sequentially calculate $t_{x_1}, t_{x_2}, \ldots$.

$$t_{x_z} = \inf\{t | \mathbb{Q}[S_m(t)] \leq G_m(t), \ \forall m\} \tag{3.13}$$

Note that in (3.13) we use $S_m(t)$ as defined in Subsection 3.2.2 and not the relaxation $S'_m(t)$.

If γ' denotes the set of implementation times associated with sequence ϑ, the reward for the sequence is,

$$r(\gamma') = \int_0^T C_t(\gamma') e^{-\alpha t} dt. \tag{3.14}$$

In summary we approximate the solution to our problem by relaxing the definition of $S_m(t)$ according to (3.9). In this way our problem has

a similar structure as a PSP. Then we use the dynamic programming formulation of Erlenkotter (3.12) to determine the optimal sequence to add patient types in the relaxed problem. From this optimal sequence we use the original definition of $S_m(t)$ and equations (3.13) and (3.14) to determine the implementation times γ' and the reward $r(\gamma')$. We call this approach the ASA.

3.4 Application and evaluation of ASA

We have introduced two approaches to solve our optimization problem (3.1). The ILP approach (described in Subsection 3.2.4) is a discrete time approximation of (3.1), although for simplicity and clarity in the text, we call the ILP solution the optimal solution. The second approach is the ASA which approximates the solution to (3.1). In this section we solve a case study problem and use numeric problem instances to evaluate and compare the two solutions approaches.

To evaluate the ASA we introduce the problem addressed at NCI. Using this problem instance as an initial case we complete a sensitivity analysis on the parameters that influence the ASA in an effort to characterize the problem instances where the ASA gives similar results as the ILP approach.

3.4.1 Application

As with many Dutch hospitals, NCI is eager to expand and provide state-of-the-art treatments with state-of-the-art equipment. To finance such expansions, managers have identified which DRGs are of the most value and hence should be the focus of the expansion. To achieve growth in the desired DRGs, the hospital must decide which patient types are the best to attract to the hospital and when to do so. In this section we apply our model to support this decision for the surgical specialty.

Since typically each patient type has a limited number of treatment options, a hospital can determine which additional patient types to accept by solving a number of subproblems. For example, historical records for the hospital in this study indicate that there are 37 different patient types (defined by their cancer diagnosis) and 109 different treatments. However some patient types have 0% probability of needing certain treatments and as such, it is possible to divide this 37 by 109 dimension problem into a number of subproblems of much smaller dimensions.

The subproblem solved in this section is called the "surgical specialty" subproblem; it encompasses all surgical treatments and their corresponding patient types resulting in six patient types and 34 treatment types. Seven of the 34 treatment types represent over 90% of the total volume of treatments, thus we only include these seven in the analysis.

Management of the surgical specialty do not expect new patient types to be added to the hospital, rather they expect the volume of existing patient types will increase if/when they try to attract them. This is essentially the same problem whereby the increased arrival rate of a patient type resulting from some hospital action, represents a new patient type. Our goal is to determine the time when this "new patient type" should be accepted.

Management also have forecasted estimates of the amount by which the volume of each patient type will increase, which corresponds to the arrival rates of the "new patient types". All data, other than these estimates, comes from the hospital's historic records. All the patient types discussed in the remainder of this section represent "new patient types".

The ILP was solved using ILOG CPLEX Solver 12.2 and the parameters (generated via the queueing model) were computed using Microsoft Visual Basic. The dynamic program used in the ASA was solved with Dijkstra's algorithm programmed in MatLAB. The parameters for the ASA were also computed using Microsoft Visual Basic.

Application data: The fraction of patients of type n that will receive

Patient Types(n)		Treatments (m) 1	2	3	4	5	6	7
1	Lung cancer	25%	0%	57%	0%	4%	0%	7%
2	Lung cancer(pleura)	22%	0%	53%	0%	8%	0%	5%
3	Breast cancer	55%	36%	0%	7%	3%	3%	0%
4	Colon cancer	24%	0%	46%	0%	9%	0%	12%
5	Colon cancer(sigmoid)	32%	0%	38%	0%	9%	0%	23%
6	Colon cancer(rectum)	4%	0%	56%	0%	0%	0%	16%

Table 3.1 – The fraction of patients of type n that receive treatment type m ($p_{n,m}$)

n	Years (20XX) 09	10	11	12	13	14	15	16	17	18	19	20
1	2.2	2.5	2.9	3.4	3.9	4.5	5.2	6.0	7.0	8.0	9.3	10
2	0.6	0.7	0.8	0.9	1.1	1.2	1.4	1.6	1.9	2.2	2.5	2.9
3	34	38	42	46	50	55	61	67	74	81	89	98
4	4.8	5.5	6.4	7.4	8.6	9.9	11	13	15	18	20	24
5	2.1	2.4	2.7	3.1	3.5	4.0	4.7	5.3	6.0	6.8	7.8	8.9
6	2.4	2.7	3.1	3.6	4.1	4.6	5.3	6.0	6.8	7.8	8.9	10
7	0.7	0.9	1.0	1.2	1.3	1.5	1.8	2.1	2.4	2.7	3.2	3.7

Table 3.2 – The mean arrival rate (patients/month) for patient type n

treatment type m is available in Table 3.1. Note that since it is possible for a patient to have multiple treatments, the sum of the probabilities across each row can be greater than 100%. In a similar manner, the sum can be less than 100% since only 7 of the 34 treatment types are included.

The relative importance of each DRG (w_m) relates to the renumeration amount that the hospital receives for each. The actual values are excluded for proprietary reasons, although they rank as follows, $w_7 > w_6 > w_4 > w_3 > w_2 > w_5 > w_1$. The forecasted arrival rate (patients / month) for each patient type n is shown in Table 3.2.

The mean service time ($\mathbb{E}[B_m]$) for each treatment type m is shown

	Treatments (m)						
	1	2	3	4	5	6	7
$\mathbb{E}[B_m]$ (months)	16.3	11.8	16.7	3.6	7.7	3.0	5.0

Table 3.3 – The mean service time for each treatment type m

	Years					
m	2009	2010	2011	2012	2013	2014
1	2004	2204	2424	2546	2673	2807
2	1573	1730	1903	1998	2098	2203
3	409	450	495	519	545	573
4	266	292	322	338	355	372
5	210	231	254	267	280	294
6	89	98	108	113	119	125
7	79	87	96	101	106	111

	Years					
m	2015	2016	2017	2018	2019	2020
1	2947	3094	3249	3411	3582	3761
2	2313	2429	2550	2678	2812	2952
3	601	631	663	696	731	767
4	391	411	431	453	475	499
5	309	325	341	358	376	395
6	131	138	145	152	159	167
7	117	122	129	135	142	149

Table 3.4 – The capacity ($G_m(t)$) to treat patients of type m over the 11 year planning horizon

in Table 3.3. The capacity ($G_m(t)$) to provide treatment type m over the 11 year planning horizon is shown in Table 3.4.

For φ_m, several values ranging from 0.1 to 0.9 were considered in order to illustrate the sensitivity of this parameter. We discretized the problem into monthly periods, meaning a patient type can only be added at the beginning of the month.

	t_1	t_2	t_3	t_4	t_5	t_6	Reward
ILP solution	0	0	53	143	100	112	24,121

Table 3.5 – Initial solution to case study (given in months)

Results: Solving (3.1) with the above data as input, the optimal reward and timing for accepting the patient types was determined and is displayed in Table 3.5. However after comparing the volume of patients receiving treatment with the available capacity, it was observed that for all treatment types (other than $m = 7$), the volume receiving treatment was much less than the capacity. It followed that treatment type $m = 7$ was the bottleneck resource and that it was significantly debilitating to the system. For example, $\mathbb{Q}[S_7(t)] > G_7(t)$ (for all t) whereas the opposite was true for the 6 other treatment types. To improve on this, we use the model to investigate how the reward changes as the capacity to provide treatment $m = 7$ is increased.

We increase the capacity of treatment type $m = 7$ by 10, 20, 30 and 40% and observed the corresponding reward. As illustrated in Figure 3.4, increasing the capacity by 30% allows the reward to increase by more than a factor of 2. Also observable in Figure 3.4, is that increasing the capacity by more than 30% does not increase the reward. Hence, a new bottleneck emerges.

Figure 3.4 demonstrates that if additional resources were made available for treatment type $m = 7$ then the hospital's reward would increase significantly. The decision is left with management whether the increase in the reward justifies the additional investment in capacity. Repeating this process of identifying the bottleneck and deciding whether additional expansions are warranted can help balance the supply and demand in the hospital. In the same vein the model can be used to determine the extent to which capacity can be decreased for certain treatment types.

After several iterations of the model, appropriate resource levels were determined and a final solution was found. The "ILP solution" row of

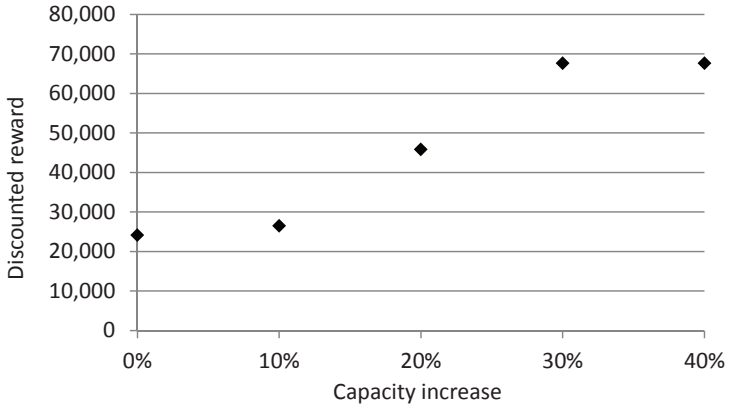


Figure 3.4 – Reward associated with increasing capacity for treatment type $m = 7$

Table 3.6 lists the reward and time when each patient type should be accepted.

Using this information the hospital can develop a strategy for enticing patients to the hospital. From the results displayed in Table 3.6, we conclude that in the short term (0 to 15 months) the hospital should focus on attracting patient types $2, 5$ and 6. In the medium term (15 to 30 months) the hospital should focus on patient types 1 and 4. In the long term, patient type 3 should be the focus. That said, the model can be used to reevaluate this policy in later years, after the level of success in attracting new patients is known. This of course could result in a new strategy to replace the current long term strategy.

3.4.2 Evaluation of approximate solution approach

The analyses of the preceding subsection were completed according to the ILP approach. In this subsection we use the same case study and compare solutions of the ASA with those from the ILP approach. Furthermore, we perform a sensitivity analysis on the main parameters

	t_1	t_2	t_3	t_4	t_5	t_6	Reward
ILP solution	25	0	53	14	0	3	67,220
ASA solution	0	0	53	53	0	11	56,529

Table 3.6 – Comparison of the ILP and ASA solutions to the case study problem instance

that impact the approximation.

To evaluate the ASA, consider that all model parameters are the same and the only difference between the ASA and the ILP approach is how we model the volume of patient receiving treatment ($S'_m(t)$ for the ASA and $S_m(t)$ for the ILP approach). Thus to evaluate the ASA, we complete a sensitivity analysis on the parameters that impact this volume ($\lambda_n(t)$ and $\mathbb{E}[B_m]$) and also the capacity parameter ($G_m(t)$). For parameters $\lambda_n(t)$ and $G_m(t)$ we vary the rate of change from t to $t+1$ and for $\mathbb{E}[B_m]$ we simply vary its value. More than 20 problem instances were evaluated to draw conclusions on the validity of the ASA. We use the following four instances to illustrate these conclusions.

Problem instance 1: This is the case study described above. Table 3.6 compares the solutions to the case study found by both approaches.

Problem instance 2: In this problem instance we increase the rate by which function $\lambda_n(t)$ changes from one period to the next. We make the rate the same for all patient types such that $\lambda_n(t+12)/\lambda_n(t) = 1.15$ for all n, meaning each patient population is increasing by 15% per year. This problem reflects a situation where the prevalence of the disease is increasing rapidly. Results for this problem instance are given in Table 3.7.

Problem instance 3: In this problem instance we decrease the rate by which function $G_m(t)$ changes from one period to the next. We make the rate the same for all treatment types such that $G_m(t+12)/G_m(t) = 1.05$ for all m. As such, this problem instance represents a hospital with modest growth ambitions. Results for this problem instance are given in Table 3.8.

	t_1	t_2	t_3	t_4	t_5	t_6	Reward
ILP solution	12	0	126	30	0	5	52,616
ASA solution	30	0	126	30	0	5	50,306

Table 3.7 – Comparison of the ILP and ASA solutions to problem instance 2

	t_1	t_2	t_3	t_4	t_5	t_6	Reward
ILP solution	37	0	79	34	16	26	20,792
ASA solution	22	16	79	33	16	37	19,954

Table 3.8 – Comparison of the ILP and ASA solutions to problem instance 3

Problem instance 4: In this example, the mean time to complete a treatment ($\mathbb{E}[B_m]$) is set to 3.5 months (for all treatments). Results for this problem instance are given in Table 3.9.

The ASA performed least well over the 4 problem instances above, for the case study. In all other instances, the ASA reward was within 6% of the ILP reward; in the case study the ASA reward was within 16% of optimal. This leads to the conclusion that good solutions can sometimes be found with the ASA. In particular, this is true when $\mathbb{E}[B_m]$ is small and when the rates of change of $\lambda_n(t)$ and $G_m(t)$ are homogeneous with respect to patient and treatment types.

However, not only is the objective function value important, but also the structure of the solution indicating when patient types should be

	t_1	t_2	t_3	t_4	t_5	t_6	Reward
ILP solution	18	0	55	26	0	16	62,704
ASA solution	55	0	55	24	0	16	60,225

Table 3.9 – Comparison of the ILP and ASA solutions to problem instance 4

accepted. In each problem instance this structure is similar for both approaches. This is particularly true in the short term (i.e. the first two or three patient types which should be accepted). In this way the ASA can be used to solve the short term plan, and then at (or near) the end of the term, the problem can be reevaluated with up-to-date parameters. As such, in the absence of an advanced ILP solver, the ASA approach can be used as the kernel of a control policy.

Furthermore, like the ILP approach, the ASA can be used to identify the bottleneck resource. For example, analyzing the case study with the ASA leads to the same conclusions about the bottleneck resource. The increase in the reward that results from adding capacity to treatment type $m = 7$ is significant (and much greater than the increase in reward which results from the using the optimal versus approximate solution). Hence improvements with respect to balancing capacity further justify the ASA.

3.5 Discussion

In this chapter we illustrate how capacity, case mix and patient mix decisions are interrelated. Understanding the complex relationships existing between these factors is crucial for achieving the maximum benefit from the DRG fee-for-service financing system. The case study highlights the impact of striving for a certain case mix without providing a balanced supply of resources.

The presented model can be extended to situations where the hospital has different motivators but still wishes to strike a balance between capacity, case mix and patient mix. In the remainder of this section we discuss three such situations.

Consider a hospital which wishes to maximize utilization of its resources instead of maximizing the number of completed treatments. By changing the reward function to account for the difference between $S_m(t)$ and $G_m(t)$ such a motivator can be accommodated.

Although this work has been applied in a hospital with fee-for-service financing, the model also has merit for health care system operating with a lump-sum budget. In such systems, hospitals typically work together with more complicated cases being treated in larger general hospitals and less complicated cases being treated in smaller regional hospitals. The decision on whether or not to treat a certain patient type at a certain hospital is similar to the problem addressed in this chapter.

The third situation considers ongoing changes in the Dutch health care system where the remuneration of some DRGs can be negotiated. In this case hospitals may strive for a certain quota of treatments, as exceeding it (or not meeting it) may result in penalties. To incorporate this, the described model can be used by making the reward depend on completed treatments per year. Furthermore, because overproduction will be penalized, the capacity constraints can be removed and the capacity levels required to meet treatment quotas become a model output.

Chapter 4

Efficiency evaluation for pooling resources

Contents

4.1 Introduction

Health care facilities are under mounting pressure to both improve quality of care and decrease costs by becoming more efficient [68, 145]. Efficiently organizing the delivery of care is one way to decrease cost and improve performance. At the national level this is achieved by aggregating services into large general hospitals in major urban centres,

thereby gaining efficiencies through economies of scale (EOS). At the same time, some hospitals are specializing and offering a limited range of services aiming to improve efficiency and decrease service rates [126]. Such strategies attempt to improve performance through focus.

At the hospital level, similar strategies to exploit focus are being considered [173, 186]. Rather than organizing departments around function (e.g. radiology, phlebotomy, etc.), departments dedicated to treating a particular patient population are being created. Examples include focused departments for back patients [203], cancer patients [124, 192], outpatients [141], trauma patients [102] and inpatients [101, 206]. In these studies, the benefits of increased focus have shown mixed results, leading to confusion over whether to become more centralized to achieve EOS or more decentralized to achieve economies of focus (EOF). In this chapter we formulate a model to measure and compare the performance of both settings. More specifically, we examine service and patient population characteristics to determine under which circumstances the centralized (functional) department, and conversely the decentralized (patient focused) department, provides better patient access times.

The chapter is organized as follows. Subection 4.1.1 introduces the principles of pooling and focus and frames the debate between centralized and decentralized departments. Using this background information, the motivation and focus of the chapter is defined in Subection 4.1.2. Section 4.2 introduces the model used to measure the EOS lost in an unpooled system. Section 4.3 describes a rough analytical approximation used to identify the main factors influencing these losses. In Section 4.4, results from simulation experiments are used to provide further perspective on these factors, to fine-tune the results and to evaluate the accuracy of the approximation. Section 4.5 summarizes the results and provides guidelines for hospital managers. In Section 4.6 we apply the model to investigate the expected service performance associated with a proposal to reallocate resources from a centralized (pooled) chemotherapy department to a breast cancer focused factory (unpooled). Section 4.7 briefly discusses potential future

research.

4.1.1 Principles of pooling and focus

The pooling principle as described in [38], is the "pooling of customer demands, along with pooling of the resources used to fill those demands [to] yield operational improvements." This implies that a centralized (pooled) clinic that serves all customer types may achieve shorter waiting times than a number of decentralized (unpooled) clinics focusing on a more limited range of customer types. The intuition for this principle is as follows. Consider the situation in the unpooled setting, when a customer is waiting in one queue while a server for a different queue is free. Had the system been pooled in this situation, the waiting customer could have been served by the idle server, and thus experienced a shorter waiting time. The gain in efficiency is a form of EOS.

Statistically, the advantage of pooling is credited to the reduction in variability due to the portfolio effect [95]. This is easily demonstrated for cases where the characteristics of the unpooled services are identical. For this discussion see [7, 62, 65, 109]. However, pooling is not always of benefit. There may be situations where the pooling of customers actually adds variability to the system, thus offsetting any efficiency gains [63]. Furthermore when the target performances of customer types differ, it may be more efficient to use dedicated capacity (i.e. unpooled capacity), see [21, 109]. And finally, in the pooled case all servers must be able to accommodate all demand. This flexibility may be expensive and, as is more directly related to this chapter, may actually cause inefficiencies as servers are no longer able to focus on a single customer type.

The principle of focus advocates for departments to limit the range of services they offer to reduce complexity and allow the department to concentrate on doing fewer things more efficiently. This philosophy has been the basis for modern manufacturing plants, which are often referred to as focused factories. Skinner [175] argues that focus, sim-

plicity and repetition in manufacturing breeds competence. The gain in efficiency due to focus is referred to in this chapter as EOF.

To exploit the principle of focus in health care, it is suggested that hospitals aggregate patients with similar diagnoses together into dedicated departments [102]. Supporters advocate "for hospitals to abandon functional, discipline-focused departments (e.g. radiology, nursing, etc.) in favor of a design organized around patients and their diagnoses" [102, 121, 153]. For example the principle of focus recommends that hospitals eliminate a centralized phlebotomy department and instead have phlebotomy services located in or near diagnosis based care department. Locating all the patient services in one department or area reduces the complexity of the process and allows care givers to oversee the complete care process from start to finish.

It is clear that pooling is offered as a potential method to improve a system's performance without adding additional resources. Interestingly, the principle of focus implies the same. In this chapter we explore these seemingly contradictory points-of-view in health care.

Other service industries have considered whether (or to what extent) resources should be pooled. In [63], the authors show that general perceptions regarding the benefits of pooling in call centres may not be in line with results from queueing theory literature. A number of practical and theoretical scenarios encountered in call centres are considered and compared numerically in [65]. Pooling of resources in the courier industry is considered in [7] where the authors use Brownian approximation models to contrast approaches used by two competing firms to provide regular and express courier services. Pooling has been studied outside of the practical domain to obtain general results. In [133], the authors consider stations in a Jackson network of queues and encourages practitioners to take care when making pooling decisions as the effect can be unbounded. In [202], the author uses approximations for $M/G/s$ queueing systems to compare various splits of pooled systems. For detailed reviews of pooling literature see [7, 63, 133].

Pooling resources to serve homogeneous demand is the common ex-

ample used to illustrate the benefits of pooling. In practice however, demand tends to be heterogeneous, in which case, these benefits are not guaranteed. Further complicating the study of the pooling of heterogeneous demands is that models tend to be analytically intractable and therefore approximate analysis is the norm [7]. Finally, most models consider continuous systems and, as discussed in Section 4.1.2, the clinics studied in this chapter do not operate in continuous time.

In this chapter and in general, the terms *pooled* and *centralized* are analogous when describing the makeup of a department or clinic. In the same way, the terms *unpooled*, *decentralized* and *focused* are analogous for describing the opposite.

4.1.2 Motivation and scope

NCI is considering the use of focused factories to treat patients with similar diagnoses. From a patient satisfaction perspective this setup is preferred, and hospital managers want to know whether service time improvements in the unpooled setting are sufficient to compensate for any EOS losses caused by unpooling the functional departments. Numerous clinic attributes influence the losses (or gains) resulting from unpooling, such as appointment length, clinic load, number of rooms, patient demand, etc. Furthermore, many of these attributes are interrelated.

When comparing the pooled and unpooled clinic makeups, the amount of resources available in each setting is the same. As such, the cost of both settings is approximately equal and our analysis compares the efficient use of resources. Comparing the efficiency of the two clinic makeups requires a definition of efficiency. In this chapter, to be consistent with the goals and constraints of the proposed focused factories at NCI, access time is the main measure of efficiency. Access time is influenced by two things, the arrival rate of new patients and the throughput of the clinic. Naturally, the arrival rate is assumed to be the same regardless of the clinic makeup. However, the throughput of patients depends on the clinic makeup. Focused clinics

are more specialized with standard practices, specialized equipment, etc., typically leading to shorter and less variable appointment durations. However, they are smaller and lack the EOS of their pooled counterpart. The analytical and simulation models described in this chapter evaluate the efficiency of both clinic makeups while reflecting the different throughput expected from each.

Specifically, the models approximate the appointment length for the unpooled system that achieves the same access time as in the equivalent pooled system. This improved service time represents the amount of improvement due to focus (or EOF) necessary to offset the losses of EOS. The approximation, along with simulations of typical clinic environments, provides the insight from which we develop general management guidelines.

The model and framework can represent any hospital department where the service time is less than one day and where the system empties between days. This includes outpatient clinics, diagnostic clinics and operating theaters. Since these departments empty at night, continuous time queueing models, which are typically used to study the effects of pooling, are not appropriate. In place of a continuous time model, a discrete time slotted queueing model is used. To the best of our knowledge such a robust model for measuring the effects of pooling and unpooling has not been developed before.

4.2 Model description

A discrete time slotted queueing model is used to evaluate the tradeoff between EOS and EOF. We describe the queueing model using language from an ambulatory clinic setting. For example, referrals for appointments are considered new arrivals, appointment length is the service time, the number of consultation rooms reflects the number of servers and finally, the time a patient must wait for a clinic appointment (often referred to as access time in health care literature) is the waiting time in the queue. In this chapter we use the following

notation,

λ = Average demand for appointments per day
D = Average appointment length in minutes
V = Variance of the appointment length
C = Coefficient of variation for the appointment length where $C = \sqrt{V/D^2}$
M = Number of rooms
ρ = Utilization of the rooms
t = Working minutes per day
W = Expected waiting time in days

A subscript "AB" corresponds to the pooled case and a subscript "A" or "B" corresponds to the unpooled case for patient groups "A" or "B" respectively. The schemes of the pooled and unpooled systems are shown in Figure 4.1.

When combined, the parameters of the unpooled system must equal the parameters of the pooled system. The parameters of the two patient groups describe the patient mix. How the patient mix parameters in the unpooled system relate to the parameters in the pooled system is described below,

$$M_{AB} = M_A + M_B \tag{4.1}$$
$$\lambda_{AB} = \lambda_A + \lambda_B \tag{4.2}$$
$$D_{AB} = qD_A + (1-q)D_B \tag{4.3}$$
$$V_{AB} = q(V_A + D_A^2) + (1-q)(V_B + D_B^2) - D_{AB}^2 \tag{4.4}$$

where $q = \lambda_A/\lambda_{AB}$.

These division "rules" imply that no additional resources become available in the unpooled setting and that patients are strictly divided into one or the other group. Note that (4.3) implies that there are no EOF gains in the unpooled setting. To incorporate EOF gains, parameter Z is introduced in Subsection 4.2.3. Although we limit our analysis to splitting a department into two groups, the results are general. This is true since splitting a department into more than two groups,

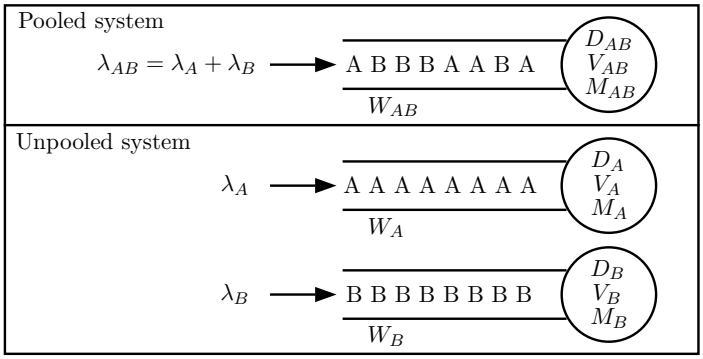


Figure 4.1 – Schematic of the pooled and unpooled systems

can be seen as splitting the original department into two groups, then splitting the resulting groups into two additional groups and so on.

Initially, the waiting times in the three queueing systems depicted in Figure 4.1 are evaluated separately. The structure of the three systems is the same and as such the same model is used to evaluate them (the input parameters are changed to reflect the pooled and unpooled systems). The approach used to evaluate the waiting times is described in Subsections 4.2.1 and 4.2.2, where the subscripts "A", "B" and "AB" are left out for clarity. In Subsection 4.2.3 we introduce a metric to compare the waiting time of the pooled and unpooled systems.

4.2.1 Modelling arrivals, services, and workload

The mean (D) and variance (V) of appointment lengths is readily available in most ambulatory clinics. Relying only on these data, we use renewal theory approximations to estimate the number of appointments completed during one clinic day. Let $N(t)$ be the number of appointments that fit into the schedule of one room between $[0, t]$. In fact, $N(t)$ is a renewal process with interarrival times distributed as appointment lengths. Further, let M be the number of rooms and

$N_i(t)$ the number of completed appointment in room $i = 1, ..., M$. We assume that $N_i(t)$s are independent and distributed according to $N(t)$. Let S be the total number of completed appointments per clinic day for a clinic with M rooms. Then,

$$S = \sum_{i=1}^{M} N_i(t). \tag{4.5}$$

We assume that the number of arrivals per day (X) is Poisson distributed with parameter λ. Then $\mathbb{E}[X] = \lambda$, $V_X = \lambda$ and $C_X^2 = 1/\lambda$, where V_X and C_X, are, respectively, the variance and the coefficient of variation of X.

Under the assumptions above the workload of the clinic (ρ) is computed by $\rho = \lambda/\mathbb{E}[S] = \lambda/(M\mathbb{E}[N(t)])$.

4.2.2 Waiting times

With the input parameters described above, our system is a *single* server system where the department as a whole is considered the server with capacity determined according to S. As such the expected queue length can be computed using Lindley's recursion [46]. Consider subsequent days $1, 2, ...$, and let L_n be the queue length at the beginning of day n. Further, let X_n be the number of arrivals on day n, and S_n the number of services that can be completed on day n. We assume that X_n and S_n, $n > 1$, are independent and distributed as described above. The number of appointment requests on day n is then $L_n + X_n$, and the dynamics of the queue length process is given by,

$$L_{n+1} = (L_n + X_n - S_n)^+; \quad n > 1 \tag{4.6}$$

where $x^+ = x$ if $x \geq 0$ and $x^+ = 0$ otherwise. When $\mathbb{E}[X_n] < \mathbb{E}[S_n]$ then for $n \to \infty$ the expectation of L_n converges to equilibrium, denoted by L [46].

To compute the expected waiting time W we use Little's Law ($W = L/\lambda$). A related model described in [192] explains how to compute the waiting time distribution through a similar recursion.

In general, (4.6) is hard to solve analytically. A variety of techniques, such as Wiener-Hopf factorization, have been developed but they usually lead to explicit solutions only in special cases. In Section 4.3 we provide a rough two-moment approximation for the average waiting time (see (4.15)). In the experiments of Section 4.4 we compute the average waiting time with simulations.

4.2.3 Required change in service time

To compare the performance of the pooled and unpooled systems, we wish to determine a new appointment length (D'_A) required to make $W_A = W_{AB}$. As a standard measure we define Z_A as the proportional difference between D_A and D'_A (likewise for D'_B and Z_B). Ignoring the subscripts "A" and "B" we formally define Z as follows,

$$Z = \frac{D'}{D} - 1. \tag{4.7}$$

Parameter Z essentially measures the EOF needed to make the access time in the pooled and unpooled systems equal. Z can be both negative and positive. When Z is negative it represents the amount the appointment length must decrease (attributed to the increased focus on a single patient group) in order to overcome any EOS losses resulting from unpooling. When Z is positive it indicates that the appointment length can increase and still maintain the same service level as in the pooled system. This happens when the number of rooms assigned to one of the patient classes is disproportionately large. Although practically less relevant, the positive Z value does help illustrate how the tradeoff between EOS and EOF is influenced by the distribution of rooms.

The convenience of metric Z is that the pooled and unpooled system can be compared without any additional input. Furthermore stake-

holders can easily interpret its meaning and decide if it is possible to obtain the necessary EOF to justify changing to an unpooled setup. In the simulation experiments of Section 4.4, Z values are computed numerically. In order to identify the system parameters that affect Z the most, in the next section we carry out a crude analysis to obtain a simple two-moment approximation (4.17) for Z.

4.3 Approximation

As Z_A depends on (4.6), which can only be obtained analytically in very special cases, we apply a simple two-moment approximation to get a rough idea about the influence of various system parameters on Z_A.

4.3.1 Two-moment approximation

To obtain the approximation formula for Z_A, we use asymptotic results from renewal theory, and thus we must assume that the appointment length is much shorter than the clinic day, i.e. $D << t$. Further, $N(t)$ in our model is the number of events on $[0, t]$ when times between events are independent and identically distributed (IID) appointment lengths. Thus, by definition, $N(t)$ is a renewal process, and with $D << t$ it follows from renewal theory [185] (pg 315) that,

$$\mathbb{E}[N(t)] \approx \frac{t}{D} + \frac{1}{2}(C^2 - 1). \tag{4.8}$$

Here, obviously, t/D is the main term, and the last term is a correction. Note from the correction term it can be observed that an increase in variability allows more events in period $[0, t]$, i.e. by increasing the variance in (4.8) the average number of completed appointments likewise increases.

Now, for the total possible number S of completed appointments, using

(4.5) we obtain,

$$\mathbb{E}[S] \approx M\mathbb{E}[N(t)] \approx \frac{Mt}{D} + \frac{M}{2}(C^2 - 1). \tag{4.9}$$

Let $V_{N(t)}$ and V_S be the variance of $N(t)$ and S respectively. From [185], the two-moment renewal theory approximation for $V_{N(t)}$ and V_S are as follows,

$$V_{N(t)} \approx \frac{V^2 t}{D^3} = \frac{C^2 t}{D} \tag{4.10}$$

$$V_S \approx MV_{N(t)} = \frac{MC^2 t}{D}. \tag{4.11}$$

We note that (4.8), (4.9), (4.10) and (4.11) are based on the assumption $D << t$. In a contrary situation (e.g. chemotherapy, where appointments may last half the day), the influence of D, V, C on S is not so direct and the above approximations cannot be used, but the general model is still valid [192].

Using (4.9) we approximate the room utilization ρ as follows,

$$\rho \approx \frac{\lambda}{\frac{Mt}{D} + \frac{M}{2}(C^2 - 1)} = \frac{\lambda D}{Mt} \frac{1}{1 + \frac{D}{2t}(C^2 - 1)}. \tag{4.12}$$

From (4.12) we observe $1/(1 + \frac{D}{2t}(C^2 - 1)) \approx 1$ when $D << t$, which is true in our case. From this observation we introduce ρ_0 as an estimate of ρ and define it as follows,

$$\rho_0 = \frac{\lambda D}{Mt}. \tag{4.13}$$

The average queue length (L) in our slotted queueing model is analogous to the average waiting time of a GI/GI/1 queue because both are measured by Lindley's recursion. In particular (4.6) corresponds to a GI/GI/1 queue with Poisson distributed service times and interarrival times distributed as S in (4.5). The waiting time of a GI/GI/1 queue can be approximated with the Allen-Cunneen approximation [3] thus

leading to an approximation for L in our slotted model. Using (4.9) and (4.11) we obtain $C_S^2 = V_S/(\mathbb{E}[S])^2$ and write the approximation formula for L as,

$$\begin{aligned} L &\approx \lambda \frac{\rho}{1-\rho} \frac{C_S^2 + (1/\lambda)^2}{2} \\ &= \lambda \frac{\rho}{2(1-\rho)} \left(\frac{1}{\lambda} + \frac{MC^2 t}{D} \frac{1}{M^2 \left(\frac{t}{D} + \frac{1}{2}(C^2 - 1) \right)^2} \right) \\ &\approx \frac{\rho}{2(1-\rho)} \left(1 + \frac{C^2}{\rho_0} \right). \end{aligned} \tag{4.14}$$

Using Little's Law and (4.14) we approximate the expected waiting time by,

$$W \approx \frac{\rho}{2(1-\rho)\lambda} \left(1 + \frac{C^2}{\rho_0} \right). \tag{4.15}$$

If λ grows and ρ remains the same then we observe a decreasing waiting time, which is credited to the EOS. Indeed, if $\lambda \to \infty$, then proportional capacity growth results in $W = 0$, see e.g. [105] for the asymptotic analysis of a similar slotted model with S equal to a constant.

Using our estimation (4.15) for W, we can also estimate the Z values based on (4.7). First we assume $\rho_0 \approx \rho$ and define ρ_0' as the load in the unpooled clinic A with appointment length D_A'. Formally we define ρ_0' as follows,

$$\rho_0' = \frac{\lambda_A D_A'}{M_A t}.$$

Next we set the waiting time approximations (4.15) for the pooled and unpooled system A equal to each other,

$$\frac{\rho_0'}{2(1-\rho_0')\lambda_A} \left(1 + \frac{C_A^2}{\rho_0'} \right) = \frac{\rho_0}{2(1-\rho_0)\lambda_{AB}} \left(1 + \frac{C_{AB}^2}{\rho_0} \right). \tag{4.16}$$

We also assume the servers are divided between the pooled and unpooled clinics in such a way that the clinic load (utilization) remains

the same. The load in the two clinics may not be exactly equal since M_{AB} and M_A must be integers. From this it follows,

$$\rho_0 = \frac{D_{AB}\lambda_{AB}}{M_{AB}t} \approx \frac{D_A\lambda_A}{M_A t}.$$

Finally, with algebra and by ignoring second order and higher terms of $(1-\rho_0)$ we solve (4.16) for D'_A/D_A to obtain,

$$Z_A = \frac{D'_A}{D_A} - 1 \approx \left(1 - \frac{1+C_A^2}{1+C_{AB}^2}\frac{\lambda_{AB}}{\lambda_A}\right)(1-\rho_0). \tag{4.17}$$

Similarly (4.17) can be rewritten to obtain $Z_B = D'_B/D_B - 1$. Using (4.4) it can be shown that either Z_A or Z_B in (4.17) is negative. In other words, when splitting a pooled department the service time of at least one of the unpooled departments must decrease to maintain the same performance level.

While deriving formula (4.17) we made a number of simplifying assumptions and ignored second order and higher terms of $(1-\rho_0)$ and the first order and higher terms of D/t . Thus, one can expect that (4.17) gives an accurate approximation for Z_A only in some special cases, e.g. when ρ_0 is close to one.

However, the main goal of deriving this formula is to reveal the main parameters that influence Z_A and to identify the relative importance of these parameters in reasonable hospital settings. To this end, our calculations show that ρ_0, λ_A/λ_{AB}, and $(1+C_A^2)/(1+C_{AB}^2)$ are the most influential factors. Furthermore, the absences of M_{AB} and D_{AB} in (4.17) implies that their influence is minimal. This is also confirmed by simulation experiments in Section 4.4. Thus, in the rest of the chapter we focus on the most influential factors appearing in (4.17).

4.3.2 Approximation results for Z_A

To illustrate the relative importance of terms ρ_0,λ_A/λ_{AB}, and $(1+C_A^2)/(1+C_{AB}^2)$ in (4.17), consider the following ranges for each of them:

	Clinic description	ρ_0	$\frac{\lambda_A}{\lambda_{AB}}$	$\frac{1+C_A^2}{1+C_{AB}^2}$	Z_A
1	Busy clinic, $\lambda_A >> \lambda_B$, $V_A << V_B$	0.99	0.7	0.32	0
2	Busy clinic, $\lambda_A >> \lambda_B$, $V_A = V_B$	0.99	0.7	1	-0.01
3	Busy clinic, $\lambda_A >> \lambda_B$, $V_A >> V_B$	0.99	0.7	1.36	-0.01
4	Busy clinic, $\lambda_A << \lambda_B$, $V_A << V_B$	0.99	0.3	0.17	0
5	Busy clinic, $\lambda_A << \lambda_B$, $V_A = V_B$	0.99	0.3	1	-0.03
6	Busy clinic, $\lambda_A << \lambda_B$, $V_A >> V_B$	0.99	0.3	2.58	-0.08
7	Quiet clinic, $\lambda_A >> \lambda_B$, $V_A << V_B$	0.7	0.7	0.32	0.16
8	Quiet clinic, $\lambda_A >> \lambda_B$, $V_A = V_B$	0.7	0.7	1	-0.13
9	Quiet clinic, $\lambda_A >> \lambda_B$, $V_A >> V_B$	0.7	0.7	1.36	-0.29
10	Quiet clinic, $\lambda_A << \lambda_B$, $V_A << V_B$	0.7	0.3	0.17	0.13
11	Quiet clinic, $\lambda_A << \lambda_B$, $V_A = V_B$	0.7	0.3	1	-0.7
12	Quiet clinic, $\lambda_A << \lambda_B$, $V_A >> V_B$	0.7	0.3	2.58	-2.28

Table 4.1 – Relative importance of factors influencing Z_A, according to (4.17)

$\rho_0 \in [0.7, 0.99]$, a higher value of ρ_0 indicates an unstable system and a lower value implies an underutilized department; $\lambda_A/\lambda_{AB} \in [0.3, 0.7]$, as having values outside of this range implies a very small unpooled department which would be impractical [192]; $C_A^2, C_B^2 \in [0.5, 3]$. Note also that $(1 + C_A^2)/(1 + C_{AB}^2)$ depends on λ_A/λ_{AB} through (4.4). Table 4.1 shows twelve scenarios reflecting the border values of these three factors.

We clearly observe that when ρ_0 is large it dominates Z_A and appears to be the most influential factor. It follows that the busier the clinic is, the smaller the loss in EOS. This is consistent with [65], who states that "pooling is not so much about pooling capacity but about pooling idleness" implying that unpooled systems with less idleness can expect less EOS gains when pooled.

Next consider that a high value of λ_A/λ_{AB} forces $(1 + C_A^2)/(1 + C_{AB}^2)$ close to 1 diminishing the effect of $(1+C_A^2)/(1+C_{AB}^2)$ on Z_A. However, for the corresponding smaller group, this factor becomes increasingly important (see rows 9 and 10 from Table 4.1).

The main goal of deriving formula (4.17) is to reveal the main parameters that influence Z and their relative importance. In the next section we use simulation to fine-tune the results for Z in a wide range of realistic scenarios. Furthermore, in Subsection 4.4.3 we evaluate the accuracy of approximation (4.17), as compared to the simulated results, for the same range of scenarios.

4.4 Numeric experiments

To gain further perspective on the factors that influence the loss in EOS and to validate the inferences drawn from (4.17) a number of simulation experiments are conducted. Subsection 4.4.1 describes the Monte Carlo Simulation and the range of the experiments. Subsection 4.4.2 provides and discusses the results of the experiments. Subsection 4.4.3 compares results of the simulation experiments with (4.17).

4.4.1 Simulation description

We model the appointment length as random variables with phase-type distributions [72,185] where expectation and variance are fitted in the data. We opt for a two moment approximation, instead of a more involved distribution fit (e.g. empirical distribution), because mean and variance data for appointment lengths are typically available. As such it is easily transferable to other settings and the likelihood of implementation is increased [192].

If the appointment length duration has $C \leq 1$ then the appointment length is assumed to follow an Erlang(k,μ) distribution where $\mu = k/D$ and k is the best integer solution to $k = D^2/V$. The completed patients per day (S) is computed by considering that an Erlang(k,μ) distribution is equal to a sum of k independent exponential random variables (phases) with parameter μ and the number of such phases

completed in t time units is Poisson with mean μt. It follows that,

$$N(t) = \lfloor \text{Poisson}(\mu t)/k \rfloor \,. \tag{4.18}$$

If $C > 1$ the appointment length is assumed to follow a hyperexponential phase type distribution. The appointment length is distributed according to $p\text{Expo}(\mu_1)+(1-p)\text{Expo}(\mu_2)$ and the total number of complete patients per day (S) is computed by Monte Carlo Simulation where,

$$p = \frac{1}{2}\left(1 + \sqrt{\frac{C^2-1}{C^2+1}}\right), \quad \mu_1 = \frac{2p}{D}, \quad \mu_2 = \frac{2(1-p)}{D}.$$

With this service rate distribution and under the assumption that the arrival rate is Poisson distributed, the waiting time and Z values are obtained by simulation. The average queue length, described by Lindley's recursion, is determined by simulating 10,000 clinic days of which 100 are used as a warm up. Little's Law is used to compute the average waiting time. To compute the Z values, the input to the simulation is systematically changed and the output compared. More specifically, Z_A is computed by incrementally decreasing [or increasing] D_A by a small amount, until $W_A \leq W_{AB}$ [$W_A \geq W_{AB}$]. The percentage change (Z_B) for patient group B is computed in the same manner. All computations are automated with Microsoft Visual Basic. Each of the simulated scenarios is described by the patient mix, clinic environment and server allotment, as introduced below.

Patient mix: The patient mix is described by two factors: λ_A/λ_{AB}, and D_A/D_{AB}. The chosen values for λ_A/λ_{AB} are 0.3, 0.4, 0.5, 0.6, and 0.7. This represents the range of situations where patient group A is 30% [group B is 70%] of the pooled group up to the situation where group A is 70% [group B is 30%] of the pooled group. The chosen values for D_A/D_{AB} are 0.5, 1, and 1.5 representing situations where the appointment length for Group A is half that of the pooled group, and up to and including the case, where it is 1.5 times longer. The appointment length of Group B can be computed easily from (4.3).

Clinic environments	M_{AB}	D_{AB}	λ_{AB}	ρ_0	C_A,C_B
1 Base clinic	20	30	282	0.88	0.5, 0.5
2 Busier clinic	20	30	*310*	*0.97*	0.5, 0.5
3 Smaller clinic	*10*	30	*141*	0.88	0.5, 0.5
4 Shorter appointments	20	*15*	*564*	0.88	0.5, 0.5
5 Higher variability	20	30	282	0.88	*2.0, 2.0*
6 Heterogeneous variability	20	30	282	0.88	*2.0,* 0.5

Table 4.2 – Parameters for different clinic environment scenarios

Clinic environment: To represent different clinic environments, the parameters for the pooled clinic are changed to represent busier clinics, smaller clinics, more variable clinics, etc. Specifically we change the values of parameters M_{AB}, D_{AB}, λ_{AB}, ρ_0, C_A and C_B. The scenarios considered are listed in Table 4.2 and are meant to encompass a wide range of typical clinic environments. The italicized values of Table 4.2 indicate the parameters which are changed relative to the Base clinic which is described in row one of Table 4.2.

Server allotment: As discussed in Section 4.2 we wish to have the same total number of servers (rooms) in the unpooled system as in the initial pooled system. The number of rooms to allot to each of the unpooled clinics needs to be decided. To illustrate how this decision impacts Z_A and Z_B consider the results in Figure 4.2 where the clinic environment is consistent with the Base clinic and the patient mix parameters are $\lambda_A/\lambda_{AB} = 0.5$, and $D_A/D_{AB} = 1$.

As illustrated in Figure 4.2, the smallest total loss in EOS corresponds with a room allotment of 10 rooms for each of the unpooled clinics. This is also the room allotment where the difference between ρ_{AB}, ρ_A and ρ_B is minimized. Let such a division be called the proportional room division, where $\rho_{AB} = \rho_A$ which implies,

$$\begin{aligned} \frac{\lambda_{AB} D_{AB}}{t M_{AB}} &= \frac{\lambda_A D_A}{t M_A} \\ M_A = \frac{\lambda_A}{\lambda_{AB}} \frac{D_A}{D_{AB}} M_{AB}, &\qquad M_B = M_{AB} - M_A. \end{aligned} \tag{4.19}$$

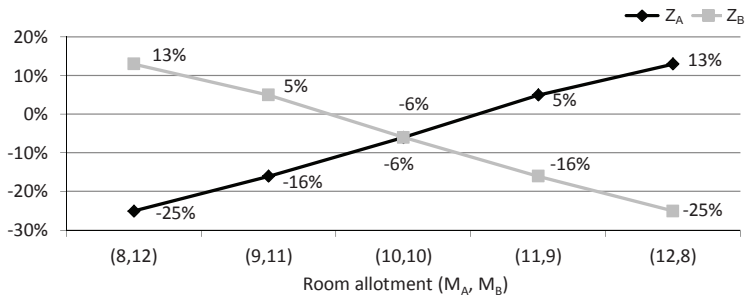


Figure 4.2 – Z values for various room allotments for the Base clinic environment where $\lambda_A/\lambda_{AB} = 0.5$, and $D_A/D_{AB} = 1$

Practically speaking this division represents the most equitable way to divide the rooms such that the difference in workload for staff in the two unpooled clinics is minimized. In the following subsections, results are only provided for the proportional room divisions.

4.4.2 Experiment results

The results in this section are organized as follows. Initially the Base clinic is analyzed for the various patient mixes. Then the clinic environment parameters are changed one-by-one and the results for each clinic environment are discussed in relation to the Base clinic.

Base clinic: The parameters and results for the initial Base clinic environment are shown in Table 4.3. The patient mix factors λ_A/λ_{AB}, and D_A/D_{AB} represent the rows and columns respectively. The results in each table cell are in the following format: $Z_A\ (M_A), Z_B\ (M_B)$. This represents the amount of change (Z_A) in D_A necessary, when the unpooled clinic is allotted M_A rooms (likewise for patient group B).

As an example consider when $\lambda_A/\lambda_{AB} = 0.3$ and $D_A/D_{AB} = 0.5$. The value in the corresponding cell is "-10%(3), -4%(17)". As noted by the numbers is parentheses, this represents the case where three rooms are allotted to Group A and 17 to Group B. In this case, for the

$\frac{\lambda_A}{\lambda_{AB}}$	$D_A/D_{AB} = 0.5$	$D_A/D_{AB} = 1.0$	$D_A/D_{AB} = 1.5$
0.3	-10%(3), -4%(17)	-12%(6), -4%(14)	-12%(9), -3%(11)
0.4	-7%(4), -5%(16)	-9%(8), -5%(12)	-9%(12), -4%(8)
0.5	-4%(5), -7%(15)	-6%(10), -6%(10)	-7%(15), -4%(5)
0.6	-3%(6), -9%(14)	-5%(12), -8%(8)	
0.7	-2%(7), -13%(13)	-4%(14), -11%(6)	

Table 4.3 – Base clinic results ($M_{AB} = 20$, $D_{AB} = 30$, $\lambda_{AB} = 282$, $C_A = C_B = 0.5$). The patient mix factors λ_A/λ_{AB}, and D_A/D_{AB} represent the rows and columns respectively. The results in each table cell are in the following format: Z_A $(M_A), Z_B$ (M_B). This represents the amount of change (Z_A) in D_A necessary, when the unpooled clinic is allotted M_A rooms (likewise for patient group B).

unpooled systems to perform equally as well as the pooled systems, Groups A and B are required to change their appointment length by $Z_A = -10\%$ and $Z_B = -4\%$ respectively. The blank cells in the table are a consequence of excluding room divisions which result in a $|Z|$ value greater than 25%.

From Table 4.3 and as identified in (4.17), Z depends on the ratio λ_A/λ_{AB}. When Group A is smaller than Group B (i.e. $\lambda_A/\lambda_{AB} < 0.5$), Group A requires less rooms but a greater decrease in service time. The counter situation (i.e. $\lambda_A/\lambda_{AB} > 0.5$) holds for Group B. It follows that larger patient groups retain EOS and require less EOF to compensate. Practically this implies that making a small department to serve a small patient population is not a good idea. This influence of λ_A/λ_{AB} is observable in all tables in this section.

Although not identified by (4.17), from Table 4.3 it appears that Z depends on the ratio D_A/D_B. This dependency is not easily characterized as it appears dependent on λ_A/λ_{AB}. Within the range of values tested, the influence of D_A/D_B is small relative to that of λ_A/λ_{AB}. This is observable in all the tables in this section except Table 4.4 where the factor ρ_0 dominates.

Busier clinic: To determine how Z_A and Z_B are influenced by how

$\frac{\lambda_A}{\lambda_{AB}}$	$D_A/D_{AB} = 0.5$	$D_A/D_{AB} = 1.0$	$D_A/D_{AB} = 1.5$
0.3	-4%(3), -3%(17)	-3%(6), -2%(14)	-6%(9), -2%(11)
0.4	-3%(4), -3%(16)	-3%(8), -2%(12)	-5%(12), -2%(8)
0.5	-3%(5), -6%(15)	-2%(10), -2%(10)	-5%(15), -3%(5)
0.6	-3%(6), -6%(14)	-2%(12), -3%(8)	-5%(18), -3%(2)
0.7	-2%(7), -9%(13)	-2%(14), -3%(6)	

Table 4.4 – Busier clinic results ($M_{AB} = 20$, $D_{AB} = 30$, $\lambda_{AB} = 310$, $C_A = C_B = 0.5$)

busy a clinic is, the demand for appointments is increased to $\lambda_{AB} = 310$. Comparing Table 4.3 with Table 4.4 it is clear that $|Z_A| + |Z_B|$ decreases as the clinic load increases. This means, that the EOS loss of unpooling is smaller for clinics of higher load. This is consistent with the findings from (4.17).

Smaller clinic and shorter appointment lengths: As expected from (4.17), the results for the clinic with fewer rooms showed only modest changes in Z_A and Z_B and are therefore excluded from the text. However, it is important to note that in smaller pooled clinics, it is less likely that (4.19) will result in a near integer solution, hence there is a discretization effect. In (4.17) we assume $\rho_{0,AB} = \rho_{0,A}$ and overlook this influence. The tests for a clinic with shorter appointments found Z_A and Z_B to also be insensitive to D_{AB} which is again what is expected from (4.17).

Higher appointments length variability: Results for a clinic with higher appointments length variability are available in Table 4.5. Relative to the Base clinic, C_A and C_B were both increased from 0.5 to 2. Contrasting Table 4.3 and Table 4.5 it is clear that $|Z_A| + |Z_B|$ has increased considerably with C_A and C_B.

Although an increase was expected from (4.17) the extent of the increase is greater than anticipated. This leads to the conclusion that changes in C_A and C_B have a greater impact than (4.17) indicates.

This is most easily illustrated by considering the patient mix $\lambda_A/\lambda_{AB} =$

$\frac{\lambda_A}{\lambda_{AB}}$	$D_A/D_{AB} = 0.5$	$D_A/D_{AB} = 1.0$	$D_A/D_{AB} = 1.5$
0.3	-22%(3), -5%(17)	-19%(6), -6%(14)	-17%(9), -7%(11)
0.4	-18%(4), -8%(16)	-14%(8), -8%(12)	-13%(12), -11%(8)
0.5	-15%(5), -11%(15)	-10%(10), -10%(10)	-11%(15), -15%(5)
0.6	-14%(6), -14%(14)	-8%(12), -14%(8)	-9%(18), -22%(2)
0.7	-13%(7), -19%(13)	-5%(14), -18%(6)	

Table 4.5 – Higher appointment length variability results ($M_{AB} = 20$, $D_{AB} = 30$, $\lambda_{AB} = 282$, $C_A = C_B = 2$)

$\frac{\lambda_A}{\lambda_{AB}}$	$D_A/D_{AB} = 0.5$	$D_A/D_{AB} = 1.0$	$D_A/D_{AB} = 1.5$
0.3		-11%(6), 4%(14)	-14%(9), 3%(11)
0.4	-23%(4), -5%(16)	-8%(8), 3%(12)	-11%(12), 2%(8)
0.5	-5%(5), 2%(15)	-6%(10), 2%(10)	-9%(15), -2%(5)
0.6	-4%(6), -2%(14)	-4%(12), -2%(8)	-5%(18), -24%(2)
0.7	-4%(7), -5%(13)	-3%(14), -4%(6)	

Table 4.6 – Heterogeneous coefficient of variance results ($M_{AB} = 20$, $D_{AB} = 30$, $\lambda_{AB} = 282$, $C_A = 2$ $C_B = 0.5$)

0.5 and $D_A/D_{AB} = 1$ which represents the case where both patient groups have equal service rate and arrival rate parameters. In this situation, the aggregate service rate for the pooled group also have the same parameters, see (4.3) and (4.4). As such, with this patient mix, C_{AB} always equals C_A and likewise C_B. In the simulation experiment for this patient mix, $|Z_A|$ increased by 4% when C_A and C_B were increased from 0.5 to 2. Evaluating (4.17) for the same situations shows no change in $|Z_A|$, illustrating that (4.17) does not fully capture the impact of C_A on $|Z_A|$.

Heterogeneous coefficient of variance: Results for the scenario when $C_A = 2$ and $C_B = 0.5$ are shown in Table 4.6. Relative to the Base clinic Z_A decreased and, with few exceptions, Z_B increases.

		$\frac{\lambda_A}{\lambda_{AB}} = 0.3$	$\frac{\lambda_A}{\lambda_{AB}} = 0.5$	$\frac{\lambda_A}{\lambda_{AB}} = 0.7$
Clinic environments	1	-28%, (-12%)	-12%, (-6%)	-5%, (-4%)
	2	-7%, (-3%)	-3%, (-2%)	-1%, (-2%)
	3	-28%, (-12%)	-12%, (-7%)	-5%, (-4%)
	4	-28%, (-10%)	-12%, (-6%)	-5%, (-3%)
	5	-28%, (-19%)	-12%, (-10%)	-5%, (-5%)
	6	-72%, (-11%)	-16%, (-6%)	-5%, (-3%)

Table 4.7 – Comparison of analytic approximation of Z_A with simulation experiments (Simulated Z_A values appear in parentheses)

4.4.3 Comparison with analytic approximation

To evaluate the accuracy of approximation (4.17) and to determine in which situations it would provide accurate estimations for Z, we compare simulated results from this section with results computed according to (4.17). Table 4.7 lists the Z_A values for the six clinic environments as computed by the simulation described in Subsection 4.4.1 and by the approximation (the simulated Z_A values appear in parentheses). Since both the simulation and (4.17) found Z to be mostly insensitive to D_A/D_{AB}, we set $D_A/D_{AB} = 1$. Furthermore, since the purpose of this subsection is to compare the two approaches we only show the Z values for Group A. Due to the symmetry however, the Z_B values can also be derived from Table 4.7.

In the derivation of (4.17) we ignored second order and higher terms of $(1 - \rho_0)$ and therefore, as expected, (4.17) is quite accurate for larger values of ρ_0 and λ_A/λ_{AB}. This corresponds with the reasonably accurate results observed in Table 4.7 for the Busy clinic (environment 2) and cases where the group size is proportionally large. In other cases simulation is a more appropriate method, especially if C is different between the two patient groups, as in clinic environment 6.

4.4.4 Conclusions

From the analytic approximation of Z we conclude that when contemplating dividing a pooled department, managers should consider ρ, λ_A/λ_{AB}, and $(1+C_A^2)/(1+C_{AB}^2)$. The importance of all three of these factors is confirmed by the simulation experiments. In the simulation experiments we also find that Z_A and Z_B values appear slightly sensitive to the ratio D_A/D_B, although characterizing this influence is not observable from the results. Furthermore, with the simulation we identified how the division of rooms between the unpooled departments is also an important decision factor. Finally the simulation also illustrates the discretization effect that occurs in smaller clinics. Both approaches used to quantify the factors impacting the unpooling decisions illustrated that there are numerous considerations necessary and many cannot be considered in isolation. Table 4.8 summarizes these factors.

Besides mean waiting times, hospitals are also interested in waiting time norms (i.e. the percentage of patients waiting less than a given target). A recursion, similar to that of Lindley's can be formulated to determine the waiting time distribution [192]. Using this waiting time recursion (instead of the queue length recursion), the simulation experiments of this section could be repeated to determine the effects of pooling with respects to waiting time norms.

Finally, although not considered in this analysis, partial pooling of resources may be a beneficial compromise to the strict resource pooling considered in this section. Partial resources pooling would see some resources dedicated to each group and the remaining resources shared between them (see [65, 202]).

4.5 Implications for practice

In general, managers should consider the following when approaching the decision to unpool a centralized department. Under most circum-

stances access time to clinics will increase unless the service time in the unpooled department is decreased, assuming that no additional resources are made available. The amount of service time decrease needed to compensate for this performance loss depends on the characteristics of the original pooled clinic and the characteristics of the newly created unpooled clinics. The main characteristics to consider are clinic load (ρ), proportional size of the patient groups (λ_A/λ_{AB}), resource division and variability in appointment length. Table 4.8 summarizes all of the characteristics/factors considered in this chapter.

When looking at the original pooled clinic consider the following. Clinics under high load require less decrease in service time to compensate for unpooling losses. The number of rooms in a clinic does not greatly influence the needed service time change, however in smaller clinics it is more difficult to proportionally divide the rooms.

When deciding how to split the pooled clinic (which consequently defines the characteristics of the new unpooled clinics) consider the following. The smallest required decrease in service time occurs when the difference between the clinic load in the two unpooled clinics is minimized. To compute the resource allocation that corresponds to this bed division see (4.19). The smaller patient group resulting from the split will require a greater decrease in service time to compensate for unpooling losses. Finally, unpooling patient groups with highly variable appointment lengths also requires a greater decrease in service time to compensate.

For more specific results refer to the tables in Section 4.4 or apply the approach described in the same section. The approach used for developing these tables is versatile in terms of the application area and practical in that it requires only typical clinical data as input.

4.6 Application

In this section we investigate a proposal at NCI to reallocate resources from a centralized (pooled) chemotherapy day unit (CDU) to a (un-

Factors	Change in Z_A	Management Guidelines
Clinic load	Decreases as ρ_0 increases	Unpooling clinics with high load results in less EOS losses than clinics under lesser load.
Room division	Disproportionate splits increase $\lvert Z_A\rvert + \lvert Z_B\rvert$	The room allotment representing the smallest loss in EOS occurs when the difference between ρ_{AB}, ρ_A and ρ_B is minimized, see (4.19).
Clinic size	Increases (slightly) as M_{AB} decreases	EOS losses appear mostly insensitive to the size of the clinic. In smaller clinics it is more difficult to proportionally split servers.
Appointment length	Mostly insensitive	EOS losses appear to be mostly insensitive to the length of the appointment.
Homogeneous variability	Increases as C_A and C_B increase	Unpooling patient groups with highly variable appointment lengths results in larger EOS losses.
Heterogeneous variability	Decreases when $C_A < C_B$	The patient group with the smaller coefficient of variance generally experiences a smaller loss in EOS as a result of unpooling.
Proportional group size	Increases as λ_A/λ_{AB} decreases	Smaller patient groups experience a greater loss in EOS as a result of unpooling.
Proportional appointment length	Mostly insensitive to D_A/D_{AB}	EOS losses appear to be mostly insensitive to the ratio of appointment lengths.

Table 4.8 – Summary of factors effecting EOS looses due to unpooling

pooled) breast cancer focused factory. The CDU has 30 beds and operates like a typical outpatient clinic with a high rate of recurring appointments. New patients are referred to the department by an oncologist who stipulates the treatment plan, including the type, dosage and frequency of infusions.

Patients arrive for their first appointment, have a short orientation meeting with a nurse and then proceed with their first treatment. Treatment involves patients sitting or lying in a bed while receiving chemotherapy drugs intravenously. This treatment could take between 30 minutes and eight hours. While receiving the drug, patients are supervised by an oncology nurse who can generally oversee multiple patients at one time. After treatment patients visit the receptionist and plan their next appointments.

The time between treatments is set by the oncologist and can range from a single day to six weeks. These recurring visits continue until the end of the treatment plan, which can be as long as one year and may change frequently as related to the condition of the patient. In an average week the clinic sees 15 new patients and 300 return patients.

As discussed previously when restrictions are added to a planning system (such as assigning beds to a particular patient group) the performance of the system will not increase unless efficiency gains result from the focused treatment. In chemotherapy treatment there are no such efficiency gains expected as the time to give treatment is dictated by the speed at which a body can absorb the drug and not by operational considerations such as setup time.

As such, the objective of this study is to investigate and quantify losses in efficiencies that can be expected as a result of separating breast cancer chemotherapy treatment (and the associated resources) from the existing centralized department. The study also investigates how best to split the existing department and where possible, how to minimize or mitigate resulting efficiency losses. This information is essential for management to decide if the inclusion benefits of having chemotherapy service in the breast cancer focused factory outweigh

Cancer type	Appointment frequency monthly	daily	% of total	Appointment duration (mins) average	variance
Breast	761	35.5	59%	160.2	4992
Non-breast	555	24.7	41%	153.7	7804
Total	1316	60.2	100%	157.5	6162

Beds	Nurses	Setup time (mins)
30	10	30

Table 4.9 – General characteristics of the chemotherapy department

the expected decrease in performance.

4.6.1 Data

For this study seven months of data was extracted from the CDU's outpatient scheduling system. The data is typical for such a system and includes patient demographics, appointment dates, times, durations, and diagnosis codes. In our study, all diagnosis codes containing keywords "Breast" and "Mamma" were considered to be breast cancer patients. The remaining diagnosis codes were scanned manually to ensure none were overlooked. From the data, statistics were computed to provide perspective on the general characteristics of the clinic. The results summarized in Table 4.9 indicate, among other things, the average and variance of appointment durations and the proportion of appointments used by the breast cancer and the non-breast cancer populations. The period of measure incidentally represents a segment of time with a lower occupancy rate due to a temporary combination of physician change and reduced demand. The principle scheduling mechanisms are however not influenced by these somewhat lower figures.

Using data on completed appointments and the department's access

time, the arrival rate distribution is derived from the seven month data set. Despite the presence of repeat visits, it follows from the data that the total number of new and return appointments completed per day can be modelled with Poisson distribution of mean 60.2. Currently patients wait a few days for their initial appointment but have their return appointments without any delay, meaning that the patient access time is negligible. The department has excess capacity and appears to operate like a queueing system with an infinite number of servers, where a server is readily available for each arriving customer. In such a system, Poisson process of departures implies that the arrival process is also Poisson with the same parameter. Thus, we assume that the number of appointment requests per day is $Poisson(60.2)$. In later sections we assume that the arrival rate distribution remains Poisson as the mean is increased.

From the chemotherapy appointment length data, we observe that the variance is less than the mean squared for all cases. As such we approximate the distributions with an Erlang distributions (see Section 4.4.1) with parameters $k = 4, 5, 3$ and $\mu = 0.025, 0.031$ and 0.020, respectively for the existing case mix, the breast cancer group, and the non-breast cancer group.

Using these distributions the model (4.18) for $N(t)$ is verified by using the historical data to compute the actual number of patients completed per bed when the bed appeared to be operating in a saturation state, meaning that the total idle time for that bed between 09:00 and 17:00 is less than 60 minutes. This 60 minutes accounts for staff breaks, which are not scheduled. For the seven months of data available, a bed operated all day in a saturation state 500 times resulting in a mean of 2.73 patients per day and a variance of 1.12. Using a Chi-square test [139], we test how well (4.18) agrees with the historic data (where $t = 480$ minutes, $k = 4$, and $\mu = 0.025$). The Chi-square value was found to equal 0.588 which is less than the critical value at an $\alpha = 0.1$ level. Thus there is no reason to believe that the maximum number of completed patients per bed is not well modelled by (4.18).

In our case study, the queue length model (4.6) is numerically solved by

a simulation programmed in Microsoft Visual Basic. The simulation completes 10 repetitions of 10,000 clinic days (100 of which are used for a warmup period). The output from the model is the utilization, and the stationary distributions of the queue length L and waiting time W.

Additionally, we compute the probability $\mathbb{P}(W \geq r)$. Observe that out of X_n arrivals on day n, at most $(\sum_{i=n}^{n+r-1} S_i - L_n)^+$ will be served before day $n+r$ because of the first-come-first-served discipline. Thus, out of these X_n patients, exactly,

$$Z_n(r) = \left(X_n - \left(\sum_{i=n}^{n+r-1} S_i - L_n \right)^+ \right)^+ \tag{4.20}$$

patients will have to wait r days or more. Now, after simulating the run of N slots, we can estimate the distribution of the waiting time as,

$$\mathbb{P}(W \geq r) = \frac{\sum_{n=1}^{N} Z_n(r)}{\text{Total Patients}}, \qquad r \geq 1. \tag{4.21}$$

4.6.2 Results

The described model has been applied to the existing department and the two proposed departments for breast cancer and non-breast cancer patients.

The model of the current CDU at NCI with 30 beds, is depicted in Figure 4.3. A summary of the model inputs and outputs is given in Table 4.10. The output from the model indicates that the existing centralized CDU has sufficient capacity to cope with the demand and variation. The access time for the clinic and the inability to accommodate return appointments on the day requested were not available from historical data. To test the sufficiency of the model to predict these metrics, we relied on expert opinion.

Next we consider the segregation of patients between the two decentralized departments, where one department with c_{BC} beds is focused

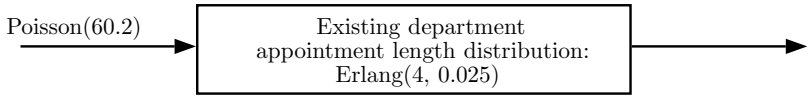


Figure 4.3 – Flow diagram of the existing centralized chemotherapy department

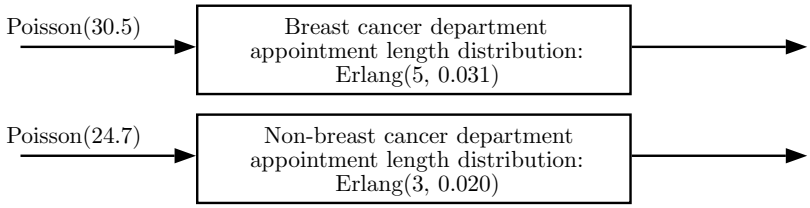


Figure 4.4 – Flow diagram of the proposed chemotherapy departments

on breast cancer patients, representing 59% of the total demand for appointments. The other department has c_{NBC} beds and is dedicated to the remaining non-breast cancer patients, which are responsible for 41% of the total demand. The two department model is depicted in Figure 4.4.

The decision on how best to distribute the 30 beds between the two decentralized departments is made according to (4.19) i.e. the proportional room division. This leads to the allotment of 18 and 12 beds to breast cancer and non-breast cancer patients respectively.

Table 4.10 summarizes the results of the performance of the existing department and the two proposed departments. As expected, the results indicate that the two decentralized departments see decreases in performance. The average queue length and waiting time increases for both. The amount that each performance metric worsens is small and seemingly unsubstantial. The main cause of this result is the excess capacity (as indicated by the low bed utilization) with which the existing department operated during the period of measurement.

The CDU currently has a bed utilization of approximately 75%. How-

	Existing dept.	Breast cancer dept.	Non-breast cancer dept.
Model inputs			
Arrival rate	Poisson(60.2)	Poisson(35.5)	Poisson(24.7)
Service time	Erlang(4,0.025)	Erlang(5,0.031)	Erlang(3,0.020)
Beds	30	18	12
Model outputs			
Service rate/bed	2.672	2.596	2.791
Bed utilization	75.1%	76.0%	73.2%
$\mathbb{E}[Lq]$	0.06	0.20	0.28
$\mathbb{E}[W]$ (days)	0.00	0.03	0.01
$\mathbb{P}(W \geq 1)$	0.1%	0.6%	1.1%
$\mathbb{P}(W \geq 2)$	0	0	0

Table 4.10 – Comparison of the performance of the existing department with the decentralized departments

ever, a recent management study set the utilization goal at 90%. To show how the decentralization decision is sensitive to expected increases in demand and bed utilization the situation was reanalyzed. The mean demand for appointments in the model was increased to 72.1 per day implying the desired 90% utilization. The result is shown in Table 4.11.

Again, all performance metrics worsen in the decentralized department. The greatest decrease in performance is the waiting time. Both decentralized departments expect around 8% of their appointments to be delayed. To achieve the same performance in the decentralized departments as in the centralized department additional resources need to be added. Table 4.11 displays the results of a bed allocation of 19 for the breast cancer department and 13 for the non-breast cancer department, which brings the total number of beds up to 32. As a result, all model outputs for performance are as good as in the existing centralized department.

As a final model output for management's consideration, the percent-

age of days that the extra bed is used in the 19/13 bed scenario (as compared to the 18/12 bed scenario) is calculated and listed in Table 4.11. In both cases this proportion is close to 30 percent of the time. Hence, it may be possible for management to allocate the 19^{th} and 13^{th} beds to the decentralized departments and only schedule and staff them 30% of the time. Since return appointments generally have a two week lead time this provides sufficient time to arrange extra staffing for the additional bed.

4.7 Discussion

The study provides both a robust model and further perspective for hospital managers who are contemplating the use of focused factories. The chemotherapy model and the general methodology has been made into a decision support tool to facilitate further analysis at NCI. Other focused factory scenarios could see patient populations divided based on blood test requirements and infusion type, dosage and frequency. Management and staff from the CDU plan to use the model to measure the impact of alternative treatment options, such as faster infusions and less frequent return appointments. Two patient demand scenarios were considered in this section, the status quo and the operational maximum given current resources. However, demand is expected to grow and, with it, the department's resources. Given patient specific demand projections, the model will be used to determine the timing and scale of capacity expansions, particularly for the breast cancer population. The methodology, which proved to be an efficient way to evaluate the impact of decentralizing services, will also be repeated for decision support for focused factory proposals in other departments.

More generally, the analytic approximation provided initial insight into the influence of the many factors causing losses in EOS, however since it is an approximation it does not fully account for them. The simulation provided more accurate results for a given range of circumstances. However, due to the large number of factors and the complex relation-

	Existing Dept.	Breast cancer dept.		Non-breast cancer dept.	
Model inputs					
Arrival rate	Poisson(72.1)	Poisson(42.6)		Poisson(29.6)	
Service time	Erlang(4,0.025)	Erlang(5,0.031)		Erlang(3,0.020)	
Beds	30	18	19	12	13
Model outputs					
Service rate/bed	2.672	2.596	2.596	2.791	2.791
Bed utilization	90.1%	91.0%	86.4%	88.6%	81.6%
$\mathbb{E}[Lq]$	2.14	3.33	1.28	2.63	0.88
$\mathbb{E}[W]$ (days)	0.03	0.08	0.03	0.09	0.03
$\mathbb{P}(W \geq 1)$	3.0%	7.8%	3.0%	8.9%	3.0%
$\mathbb{P}(W \geq 2)$	0	0	0	0	0
Proportion of days extra bed is used			33.2%		29.7%

Table 4.11 – Comparison of a highly utilized centralized department with decentralized departments

ships that exist between them, it proved difficult to use simulation to draw stringent general conclusions. Further research is required to determine how exactly these factors influence losses of EOS related to unpooling. With comprehensive descriptions of these relationships, operational researchers can further improve or even optimize the mix of the functional and patient focused departments within a hospital.

Chapter 5

Panel sizing in oncology

Contents

5.1 Introduction

As with many other chronic diseases, cancer is a disease from which patients never completely recover. This means that patients, after receiving their initial treatment, may relapse and return years later for further treatment and/or palliative treatment. This has implications for capacity planning, as the number and type of patients treated today is an indicator of future demand. Incorporating patient relapses into capacity planning is essential for ensuring capacity is sufficient to meet demand over the long term.

When a patient is diagnosed with cancer, they are referred to a cancer centre by a general practitioner (GP). The patient then meets with the oncologist and a course of treatment is determined. Treatments typically consist of surgery, chemotherapy, radiotherapy, hormone therapy, or a combination thereof. During treatment a patient is "actively followed" by an oncologist, which means they have recurring appointments with the oncologist. These appointments are used to audit treatment progress, discuss side effects and complications, and/or to evaluate psycho-social impact. The frequency of these appointments depends on the patient and the treatment, but is typically between 2 and 12 appointments per year. If the treatment is successful, the patient is discharged (i.e. is no longer actively followed by the oncologist) when the cancer is in remission. If the treatment is unsuccessful, the patient receives palliative treatment.

Those patients that receive successful treatment are discharged and are monitored by their GP to identify if a relapse occurs. If a patient relapses, they are referred back to their original oncologist for further treatment. Following this, and when the cancer is in remission, the patient is again discharged. Patients can continue to be referred, treated and discharged for as long as the patient lives, although typically patients only relapse once or twice.

Since there is a limit to the number of appointments an oncologist can offer, there is a limit to the number of patients a single oncologist should follow; consequently, there is a limit to the number of new patients an oncologist should see. If an oncologist sees too many new patients now, their clinic can become overwhelmed in the future. Given that the time between referral and discharge, the relapse rate, and the arrival rate of new patients are all uncertain, the number of new patients a single oncologist should see is not immediately obvious. In this chapter we model the demand for oncology appointments over a long term planning horizon and account for both new and relapsed patient demand. We use this model to support capacity planning decisions and to investigate policies to improve oncology clinics.

In literature the number of patients a physician should follow and be

accountable for is commonly referred to as "panel size" [86]. Typically, panel sizing models compute the overflow probability for a given panel size, where the overflow probability is the probability that a day's demand for appointments is greater than the supply of appointments. In practical terms the overflow probability represent the frequency that a clinic will run into overtime. An overflow probability of 5%, 10% and 20% equates to needing overtime approximately once per month, once per two weeks and once per week, respectively [81, 86].

For a discussion on the clinical implications of an incorrect panel size see [149, 150]. A web based application designed for physician use is available [93]. All of these studies consider a primary care practice and model the individual GPs separately. The introduction of open-access scheduling in the primary care practice is often used as a motive to study panel sizes and is reviewed in [11, 149].

In [80] the impact of patient no-shows on the ideal panel size is investigated with theoretical queueing models and a simulation model. The paper clearly illustrates that much smaller panel sizes are needed in the presences of no-shows and that results for typical primary care practices with advanced access can be reliably approximated with queueing models. In [10, 11], multiple patient types are considered to reflect the different demand patterns associated with different patient demographics. For example, female patients between age 55 and 60 will typically require more appointments than male patients between age 25 and 30 [11]. The authors use simulation to determine the number of appointments expected in a given week.

The initial studies described in [81, 86] treat the panel size as a constant and do not allow a backlog in demand to occur. The overflow probabilities are computed under the assumption that all demand for a given day would be filled on that day. A later paper [80] relaxes this assumption and allowed a finite backlog to develop. The authors assume a constant panel size and conceptualize the appointment system as a single server queuing model to track backlog and measure the performance of the clinic.

The main contribution of our work to this literature is a model allowing the panel size to be a random variable by accounting for new patients arriving, and existing patients departing. We consider both cases where this random variable is stationary in time and where it is non-stationary in time. The former reflecting an established physician who has a mature practice and the latter reflecting a physician with a new practice. In both cases we allow multiple patient types and use queueing network models to determine the random panel size. In the stationary case for the mature practice we allow and measure a backlog of patients. In the non-stationary case for a new practice that has not been fully populated yet, we assume there is excess capacity and that the physician will not allow a backlog of patients (i.e. overtime is used to fulfill all of a period's demand). The frequency of overtime is determined by the model.

The setting examined in this chapter is an oncology clinic. There are certain characteristics of oncology which, in and of themselves, constitutes an extension of the previous panel size studies of GP clinics. One distinguishing characteristic of the oncology setting is that patients can have periods where they are having recurring appointments, followed by periods where they have no appointments (e.g. when the cancer is in remission). This can repeat for as long as a patient lives. Not all cancer modalities are the same, and as such, the problem of coping with relapsing patients is less pronounced for those with lower survival rates and those with lower probabilities of relapse, e.g. lung cancer. Prostate cancer and breast cancer however have higher survival rates and hence greater relapses, and therefore are the cancer modalities to which the model described in this chapter is most directly applicable.

Like the reviewed papers, we model individual physicians separately. Pooling the physicians together may yield efficiencies however in the interest of continuity of care [121] this is often not done. As such, we assume patients are always seen by the same oncologist. To the extent that the overall capacity of a region can be seen as the sum of its oncologists, the models, aid in capacity planning decisions at the regional level as well.

Given that the panel size is a random variable dependent on the turn over of a clinic, we use the mean arrival rate of new patients as the control parameter. In our experience, this metric has more meaning to oncologists than the panel size. The performance of the clinic in the stationary setting is measured in terms of the expected waiting time and in the non-stationary setting by the frequency of overtime.

The chapter is organized as follows. In Section 5.2 the model for the oncology setting is described and the relationship between the panel size and appointment frequency is defined. In this section we describe how the appointment frequency can be computed when the panel size is a random variable. This model is general in that it is applicable for both the stationary and non-stationary settings. In Section 5.3 two queueing network models are introduced to analyze when the system is in equilibrium and when it is not (i.e. when the arrival rate is stationary and non-stationary in time). In Section 5.4 an application is discussed and numeric results are presented leading to general results on maximizing the capacity to treat new patients. Section 5.5 adds concluding remarks about the models in this chapter and discuss some areas for future research.

5.2 Model description

This section is divided into two subsections; in the first subsection we describe an oncologist's practice as a queueing network and in the second subsection we describe how to translate the patients in the system into appointment demand.

5.2.1 Patients in the system

Consider that a patient being followed by an oncologist can exist in any one of three states: 1) the new patient state, 2) the active patient state and 3) the inactive patient state. Newly referred patients immediately enter the new patient state and have an initial appointment

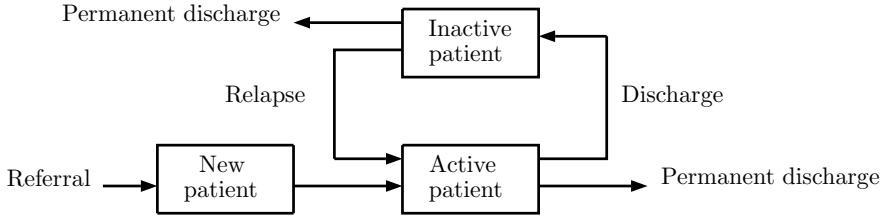


Figure 5.1 – The dynamics of active and inactive patients

with the oncologist. After this initial appointment the patient enters the active patient state and has recurring appointments with the oncologist which typically occur simultaneously with treatment (we do not model any treatments, only oncologist appointments). These recurring appointments continue until the cancer is in remission and the oncologist discharges the patients. At this point, the patient enters the inactive patient state. If an inactive patient's cancer relapses, then the patient returns to the active patient state. Finally, a patient exits the system when they die or choose to be followed by a different oncologist (for simplicity we aggregate these groups together) and can do so from either the active or inactive states. This process of patients transitioning between new, active and inactive states is illustrated in Figure 5.1.

The amount of time a patient stays in the each state, and consequently the rate at which patients switch states, depends on their treatment. Let $i = 1, 2, \ldots, c$ be different patient types reflecting the treatment a patient receives while in the active patient state. Let $N(t)$, $A_i(t)$ and $I(t)$ respectively represent the number of new, active of type i, and inactive patients at time t. And let $PS(t) = N(t)+A_1(t)+A_2(t)+\ldots+A_c(t)+I(t)$ be the panel size. The length of time a patient remains in each state has distribution D_x^i and mean $\mathbb{E}[D_x^i]$ where $x = np$ denotes the new patient state, $x = ap$ denotes the active patient state and $x = ip$ denotes the inactive patient state.

When a patient's length of stay in a state ends they are routed to another state (or out of the system) as depicted in Figure 5.1. Let

$\lambda(t)$ be the mean arrival rate at time t. All new arrivals immediately enter the new patient state. Upon exiting the new patient state all patients enters the active patient state and with probability p_i become patients of type i. When exiting the active patient state, a patient of type i enters the inactive patient state with probability $p_{i,ip}$, and with probability $1 - p_{i,ip}$ the patient exits the system. Inactive patients can relapse and with probability p'_i the relapsed patient becomes a patient of type i when they enter the active state. With probability $1 - \sum_i^c p'_i$ an inactive patient never returns to the active patient state and exists the system.

Both of the queueing models presented in Section 5.3 can be extended so that the treatment a relapsed patient receives depends on the treatment(s) they have previously received, i.e. probability p'_i can be made to depend on the type of patient he/she was in the past. In the interest of clarity, we choose the simpler formulation described above, however, when discussing the queueing models we also explain how to accommodate this extension.

5.2.2 Required appointments

In general, there are two types of appointments offered by the oncologist, "new patient" appointments and "follow-up" appointments. New patients receive new patient appointments and active patients receive follow-up appointments. Inactive patients do not receive appointments. We are interested in determining the number of new patient appointments and follow-up appointments required during the planning horizon for the clinic. For example if an oncologist wishes to offer 100 appointments per month, then the planning horizon in one month. Let δ be the planning horizon for the clinic.

To compute the required number of new patient appointments in period δ, consider that each new patient receives only one new patient appointment before entering the active patient state. Because of their urgency, new patients must have their new patient appointment within period δ. Therefore the number of new patient appointments during

period δ, equals the number of patients in the new patient state during period δ, and also equals the number of new patient arrivals in period δ. It follows, that $N(t)$ denotes the number of patients in the new patient state at the t^{th} time period of length δ, and, $N(t)$ also denotes the number of new patient appointments at the t^{th} time period of length δ, where,

$$\mathbb{E}[N(t)] = \lambda(t). \tag{5.1}$$

In the same way we are interested in the number of follow-up appointments. To compute the number of follow-up appointments required during period δ we further specify what it means to be a patient of type i. Let a patient of type i have an appointment schedule such that the time between appointments is $i \cdot \delta$ (e.g. when $i = 2$ then the patient has an appointment every second period δ while in the active patient state). This corresponds with practice in that the follow-up protocols dictate the amount of time between appointments (typically the number of months). Note that this also implies that δ is the shortest time interval between consecutive appointments for any active patient type i.

Let $F(t)$ denote the number of follow-up appointments at the t^{th} time period of length δ. Since each patient has a probability of $1/i$ of requiring an appointment in period δ then it follows that,

$$\mathbb{E}[F(t)] = \sum_{i=1}^{c} \mathbb{E}[A_i(t)]/i. \tag{5.2}$$

Let $F_i(t)$ denote the number of follow-up appointments for patient type i at the t^{th} time period of length δ. Knowing that each patient has probability $1/i$ of requiring an appointment in that period, then $F_i(t)$ can be modelled with a binomial distribution as follows,

$$\mathbb{P}(F_i(t) = k | A_i(t)) = \binom{A_i(t)}{k} \left(\frac{1}{i}\right)^k \left(1 - \frac{1}{i}\right)^{A_i(t)-k} \tag{5.3}$$

Since the probability of needing an appointment $1/i$ is different for each patient type the sum of the c random variables $F_i(t)$ (i.e. random variable $F(t)$) is not binomially distributed. However the mean is available from (5.2).

Equations (5.2) and (5.3) hold for a more general setting than oncology as well. For example, in other settings where the panel size is uncertain and where there are multiple patient types, these equations can be used.

To determine the total appointment demand we need to aggregate the new patient appointment demand with the follow-up appointment demand. Assume the total capacity of an oncologist is divided into slots and that new patient appointments require a slots and follow-up appointments require b slots. For example, if an appointment slot is 10 minutes, a new patient appointment is 60 minutes, and a follow-up appointment is 20 minutes, then $a = 60/10$ and $b = 20/10$. Let $G(t)$ be the aggregate demand for appointment slots, then,

$$\mathbb{E}[G(t)] \quad = \quad \mathbb{E}[N(t)]a + \mathbb{E}[F(t)]b. \tag{5.4}$$

With the queueing models of the next section, we derive the number of new and active patients in the system from the new patient arrival rate $\lambda(t)$. We also use the queueing models to measure the performance of the clinic associated with this new patient arrival rate. Our problem is thus: to determine the new patient arrival rate that leads to the desired appointment demand while satisfying certain performance criteria. Performance in the stationary setting is measured by the mean waiting time for active patients to receive their first appointment and in the non-stationary setting by the frequency of overtime. The motivation for these performance metrics is given in the following section.

5.3 Queueing network models

This section is divided into two subsections. The first subsection describes a multi-class open queueing network model with a stationary arrival rate. This represents an oncologist with a mature clinic (i.e. a clinic that receives, on average, the same number of new patients per month). The second subsection describes an network of infinite server queues with a non-stationary arrival rate. This queueing network represents a new oncologist's clinic.

In the stationary setting we assume a stable system and that $\lambda = \lambda(t), t \geq 0$. It follows that random variables, $PS(t)$, $N(t)$, $A_i(t)$, $I(t)$, $F(t)$ and $G(t)$ are also all stationary with respect to time. As such, for clarity, in the stationary setting these variable are respectively denoted by PS, N, A_i, I, F and G.

5.3.1 Panel size in equilibrium (stationary arrival rate)

In the stationary setting, we model each patient state as a single queue and the time a patient remains in the state (D_x^i) as the service time for that queue. We consider the network to be a multi-class network such that each patient type i is a separate patient class. The service time in the new patient queue is the same for each patient type, as such $D_{np}^i = D_{np},\ i = 1, 2, \ldots, c$. This is also true for the inactive patient queue and likewise $D_{ip}^i = D_{ip},\ i = 1, 2, \ldots, c$. In the active patient queue the service time is patient class specific.

The number of servers in the new patient queue (m_{np}) and the inactive patient queue (m_{ip}) is assumed to be infinite. For the inactive patient queue this is a natural assumption since these patients do not consume resources and arrivals are never delayed. For the new patient queue this reflects the urgency of new patient referrals, and the practice that ensures they are seen within δ, even if overtime is necessary. The number of servers in the active patient queue (m_{ap}) represents the capacity for active patients, i.e. the number of active patients that an

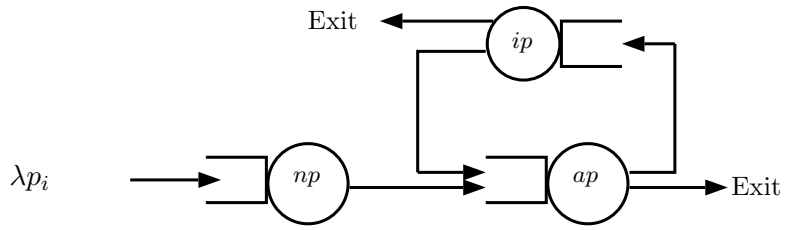


Figure 5.2 – Stationary queueing network model schematic. Queues np and ip have infinite servers and queue ap has m_{ap} servers. The time spent in each queue (waiting and in service) represents the time a patient spends in the corresponding state.

oncologist can have. Although this may not be a parameter that an oncologist can readily provide, we show how it can be determined by the model for a given arrival rate and a desired mean waiting time.

Let $\{np, ap, ip\}$ be the set of queues in the network and let the routing matrix $r^i_{k,j}$ be as follows,

$$r^i_{k,j} = \begin{cases} 1 & \text{if } k = np,\ j = ap,\ i = 1, 2, \ldots, c \\ p_{i,ip} & \text{if } k = ap,\ j = ip \\ p'_i & \text{if } k = ip,\ j = ap \\ 0 & \text{otherwise,} \end{cases}$$

with the queuing model is depicted in Figure 5.2.

The main performance metric we are interested in, is the average delay in becoming an active patient ($\mathbb{E}[W]$). Patients entering the new patient and inactive patient queues are "served" without delay. Due to the limited number of servers in the active patient queue, patients joining this queue can form a backlog and experience a delay. We assume that patients joining this queue do not preempt existing patients. Practically this implies that existing patient appointments are never delayed in order to fit in newly arriving patients, i.e. once you become an active patient, you always receive your appointments on schedule. This is reasonable, given that the appointments for existing patients are made $\delta \cdot i$ periods in advance. We are interested in knowing the

average delay in becoming an active patient as a function of the arrival rate of new patients (λ). In this stationary setting $\lambda(t) = \lambda, t \geq 0$ and the mean arrival rate of new patients of type i is λp_i.

To analyze the queueing network we use the approximate decomposition [17] method for multi-class open queueing networks. Exact methods are only possible in a few situations (e.g. Jackson networks [83] and BCMP networks [12]) which require assumptions that are too limiting for our purposes. Approximate decomposition methods are reviewed in [17] and applied in a clinical setting by [211]. For this approximation, it is assumed that arrivals are independent and identically distributed (IID) and that the first and second moments of the arrival rate and service time distributions are known. Requiring only the first and second moments for these distributions is particularly convenient as data of this sort is typically available from many hospital information systems.

The decomposition of multi-class open queueing networks is typically described with three steps. The first step aggregates the multiple patient types (classes) together, the second step analyzes the networks as if it was single class queueing network and the third step disaggregates the performance metrics to obtain measures for each individual patient type. In our network we are only concerned with the aggregate waiting time and only in the active patient queue. Therefore we do not disaggregate the waiting time to be patient type specific (i.e. we only complete steps 1 and 2). The majority of the steps used to aggregate the multiple patient classes are explained in the appendix.

We use the multi-class open queueing network model to create an explicit relationship between the mean arrival rate (λ), the mean waiting time ($\mathbb{E}[W]$), and the mean number of appointment slots ($\mathbb{E}[G]$). The first two parameters (λ and $\mathbb{E}[W]$) are model inputs and the third ($\mathbb{E}[G]$) is the model output. As such, we use an algorithmic approach to determine the mean arrival rate λ that is appropriate for a clinic. Namely, for a given $\mathbb{E}[W]$, and an initial λ, we compute $\mathbb{E}[G]$. We then compare $\mathbb{E}[G]$ with what is desired by the oncologist and adjust λ accordingly. This process is repeated until the appropriate λ value

is determined.

To determine $\mathbb{E}[G]$ for a given λ and $\mathbb{E}[W]$, we first determine the aggregate arrival rate to the active patient queue for each patient type i. Then, using Little's Law, we determine the number of active patients. Given that for the system to be stable, we must offer more appointments on average than is demanded on average, we next solve for the server utilization resulting from λ and $\mathbb{E}[W]$. Using this server utilization, we can compute the number of follow-up appointments that need to be offered and then using (5.4) we determine the total number of appointment slots.

First we determine the aggregate mean arrival rate to the active patient queue for each patient type. From [17], it follows that the aggregate mean arrival rate can be determined from the following system of linear equations,

$$\begin{aligned} \lambda^i_{np} &= \lambda p_i \\ \lambda^i_{ap} &= \lambda^i_{np} + \lambda^i_{ip} r^i_{ip,ap} \\ \lambda^i_{ip} &= \lambda^i_{ap} r^i_{ap,ip} \end{aligned}$$

which leads to,

$$\begin{aligned} \lambda^i_{np} &= \lambda p_i \\ \lambda^i_{ap} &= \frac{p_i \lambda}{1 - r^i_{ip,ap} r^i_{ap,ip}} \\ \lambda^i_{ip} &= \frac{r^i_{ap,ip} p_i \lambda}{1 - r^i_{ip,ap} r^i_{ap,ip}}. \end{aligned}$$

This queueing model can be extended so that the treatment a relapsed patient receives, depends on the treatment(s) they have previously received. This is accomplished by specifying patient types by the set of their previous treatments(s) and then making p'_i depend on that set. Should the data indicate that this is necessary, the underlying model and solution technique remain valid, however it will require solving a larger system of linear equations to determine the aggregate mean arrival rates.

We use Little's Law to compute the mean number of patients of type i that are in service at queue ap [205] as follows,

$$\mathbb{E}[A_i] = \lambda_{ap}^i \mathbb{E}[D_{ap}^i]. \tag{5.5}$$

For stability, $\sum_{i=1}^{c} \mathbb{E}[A_i] < m_{ap}$ and it follows that,

$$m_{ap} \approx \frac{\sum_{i=1}^{c} \mathbb{E}[A_i]}{\rho_{ap}}, \tag{5.6}$$

where ρ_{ap} is the utilization of the m_{ap} servers. To determine ρ_{ap}, consider that the mean waiting time in the active patient queue can be approximated using a $G/G/c$ approximation. Since this is a multi-class queueing network we first aggregate the service time and arrival process of all patient types together (this process is described in detail in the appendix). The mean waiting time can be estimated as follows (see [169]),

$$\mathbb{E}[W] \approx \frac{SCV_{ap,a} + SCV_{ap,s}}{2} \frac{\rho_{ap}^{\left(\sqrt{2(m_{ap}+1)}-1\right)}}{m_{ap}(1-\rho_{ap})} \mathbb{E}[D_{ap}^{+}], \tag{5.7}$$

where $SCV_{ap,a}$ and $SCV_{ap,s}$ are respectively the squared coefficient of variance for the aggregate arrival and service processes at queue ap and $\mathbb{E}[D_{ap}^{+}]$ is the mean aggregate service time in queue ap. Using (5.6) and (5.7) we are left with two equations and two unknowns (ρ_{ap} and m_{ap}) and thus we can solve for both.

Knowing the utilization of the servers ρ_{ap} allows us to compute the number of follow-up appointments that must be offered as follows,

$$\mathbb{E}[F] = \sum_{i=1}^{c} \frac{\mathbb{E}[A_i]}{i\rho_{ap}}$$

and the total number of appointment slots that must be offered in this stationary setting is,

$$\mathbb{E}[G] = \lambda a + \mathbb{E}[F]b.$$

Using Little's Law the mean panel size is the stationary setting is,

$$\mathbb{E}[PS] = \sum_{i=1}^{c} \left(\lambda p_i + \lambda_{ap}^i (\mathbb{E}[D_{ap}^i] + \mathbb{E}[W]) + \lambda_{ip}^i \mathbb{E}[D_{ip}] \right).$$

5.3.2 Panel size when not in equilibrium (non-stationary arrival rate)

In this subsection we consider the arrival rate of new patients to be non-stationary in time. Like in the stationary model, we model the new patient state and the inactive patient state each with an infinite server queue. Unlike the stationary model, we model the active patient state with c queues (one for each patient type) and consider only a single class of patients. Each of the c queues for the active patient state is also considered to have infinite servers. This means that all patients joining the active patient queues will receive their appointments without delay and all recurring appointments occur without delay (i.e. there is no backlog of demand).

This is perhaps an idealized case. However, it does represent the practice of oncologists who adjust their schedule (e.g. uses some non-clinic time to fit in extra patients) in order to see all patients within the required period (this is true in the oncologist clinic that motivated this study). This is particularly fitting for a new oncologist's clinic who is building up to the stationary panel size. During this time it is reasonable to assume that the oncologist has excess capacity and is able to accommodate all of the demand for period δ in period δ. Like in the models of [81, 86] the frequency of needing overtime is the metric by which the performance of the clinic is measured. A schematic of the queueing model is shown in Figure 5.3.

We use the network of infinite server queues to create an explicit relationship between the mean arrival rate ($\lambda(t)$), the frequency of overtime ($\mathbb{P}(G(t) > g)$), and the mean number of appointment slots ($\mathbb{E}[G(t)]$). Where g is the number of appointment slots offered by the oncologist in period δ. We use the model to compute the maximum

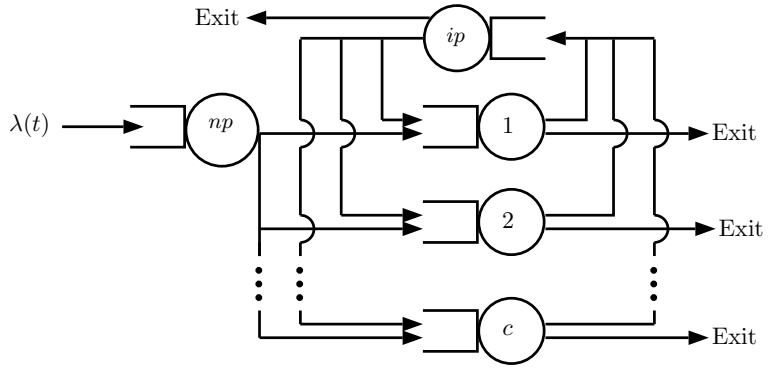


Figure 5.3 – Non-stationary queueing network model schematic. All queues have infinite servers. The time spent in service in each queue represents the time a patient spends in the corresponding state.

$\lambda(t)$ value possible such that the frequency of overtime is less than φ. Let $\lambda^*(t)$ denote this maximum. It follows that,

$$\lambda^*(t) = \min\{\bar{\lambda}(t), \max\{\lambda(t) | \mathbb{P}(G(t) > g) \leq \varphi\}\} \tag{5.8}$$

where $\bar{\lambda}(t)$ is the maximum arrival rate possible in period t based on the prevalence of cancer in the hospital's catchment area. With (5.8) we are essentially maximizing the utilization of the new oncologist while ensuring that the frequency of overtime is less than φ . Since $G(t)$ depends on $A_i(t)$, the demand of patients already in the system is accounted for.

Since the patient population is large, and patients get ill independently of each other, it is natural to assume that new patients arrive according to a Poisson process. Furthermore, since trends and seasonality are common in many diseases it is also natural to assume a non-stationary Poisson process. Let $J = \{1, 2, \ldots, c, np, ip\}$ be the set of infinite server queues in the queueing network and let j be the index of the queue in the set. Let $H^j(t)$ be the number of patient in queue j at time t and

let the service time in each queue be, D^j such that,

$$D^j = \begin{cases} D_{np} & \text{if } j = np \\ D_{ip} & \text{if } j = ip \\ D^j_{ap} & \text{if } j = 1, 2, \ldots, c. \end{cases}$$

Using the result from [27, 135] for networks of infinite server queues with non-stationary Poisson arrival processes, it can be readily seen that the number of patients in queue j at time t ($H^j(t)$) is distributed according to,

$$\mathbb{P}(H^j(t) = a) = \frac{e^{-\mathbb{E}[H^j(t)]}(\mathbb{E}[H^j(t)])^a}{a!}.$$

Note that $H^j(t)$ is a Poisson distributed random variable with mean,

$$\mathbb{E}[H^j(t)] = \mathbb{E}[\lambda_j^{'+}(t - D^{j,e})]\mathbb{E}[D^j],$$

where $D^{j,e}$ is the excess service time with the following cumulative distribution function,

$$\mathbb{P}(D^{j,e} \leq t) = \frac{1}{\mathbb{E}[D^j]} \int_0^t \left(1 - \mathbf{D}^j(u)\right) du,$$

and $\mathbf{D}^j(u)$ is the cumulative distribution function of the service time distribution. And where $\lambda_j^{'+}(t)$ is the aggregate arrival rate function to queue j and is the minimum nonnegative solution to the following system of equations,

$$\lambda_j^{'+}(t) = \lambda(t)\pi_j + \sum_{k \in J} \mathbb{E}[\lambda_k^{'+}(t - D^k)] r_{k,j},$$

and $\pi_j = 1$ when $j = np$ and $\pi_j = 0$ otherwise. Routing matrix $r_{k,j}$ is as follows,

$$r_{k,j} = \begin{cases} p_k & \text{if } k = np,\ j = 1, 2, \ldots, c \\ p_{k,ip} & \text{if } k = 1, 2, \ldots, c,\ j = ip \\ p'_k & \text{if } k = ip,\ j = 1, 2, \ldots, c \\ 0 & \text{otherwise.} \end{cases}$$

To determine the number of patients in each state note that,

$$H^j(t) = \begin{cases} N(t) & \text{if } j = np \\ I(t) & \text{if } j = ip \\ A_j(t) & \text{if } j = 1, 2, \ldots, c. \end{cases}$$

Expressions for the departure process are also available in [135]. This queueing model can also be extended so that the treatment a relapsed patient receives, depends on the treatment(s) they have previously received. This is accomplished by replacing queue ip with a multiple queues. Each of these new queues would be indexed by a set of treatment(s), and patient would be routed to the queue that corresponds to the set of treatments they have previously received. This results in a different (and larger) set of queues J (and different routing probabilities) but the underlying model and solution technique remain valid.

Since $A_i(t)$ is Poisson distributed, it follows that $F_i(t)$ is also Poisson distributed. Furthermore, since patients arrive according to a Poisson process, it follows that $N(t)$ is also Poisson distributed. Finally, since $G(t)$ is the aggregate of independent random variables $N(t), F_1(t), F_2(t), \ldots, F_c(t)$ which are all Poisson distributed, then $G(t)$ is also Poisson distributed.

Let g be the number of appointments slots offered in period δ by a new oncologist. It follows that the frequency of overtime (i.e. the probability that demand is greater than g) is computed by,

$$\mathbb{P}(G(t) > g) = 1 - \sum_{k=0}^{g} \frac{e^{\mathbb{E}[G(t)]}(\mathbb{E}[G(t)])^k}{k!}$$

where,

$$\mathbb{E}[G(t)] = \mathbb{E}[N(t)]a + (\mathbb{E}[F_1(t)] + \mathbb{E}[F_2(t)] + \ldots + \mathbb{E}[F_c(t)])\, b$$

and $\mathbb{E}[F_i(t)] = \mathbb{E}[A_i(t)]/i$. Finally the expected panel size in the non-stationary setting is,

$$\mathbb{E}[PS(t)] = \mathbb{E}[N(t)] + \mathbb{E}[A_1(t)] + \mathbb{E}[A_2(t)] + \ldots + \mathbb{E}[A_c(t)] + \mathbb{E}[I(t)].$$

5.4 Application

In this section we use numeric examples to illustrate how we compute the mean new patient arrival rate for an oncologist. The section is divided into two subsections, the first representing an established oncologist (i.e. a stationary setting) and the second representing a new oncologist (i.e. the non-stationary setting). In the first subsection the mean arrival rate per month λ is determined for a range of desired waiting times and appointment offerings. We then vary several model parameters to illustrate their sensitivity and to investigate the impact of two policy changes. In the second subsection, we investigate the process of a new oncologist becoming an established oncologist and determine the mean arrival rate $\lambda(t)$ which allows this to happen as quickly as possible. The planning horizon is for one month and the discrete steps in the model are also one month (i.e. the rate parameters of the model are per month), therefore $\delta = 1$.

To allow a comparison of the two queueing models, the analysis of Subsections 5.4.1 and 5.4.2 are done with the data from the same oncologist. Namely the patient mix parameters $(p_i, p_i', p_{i,ip})$ and the time spent by patients in each state (D_{np}, D_{ap}^i and D_{ip}) are the same in both subsections. The data comes from an oncology clinic located in Vancouver, Canada, and which is part of the British Columbia Cancer Agency (BCCA). Ten years of data extracted from the hospital's information management system was used for this analysis.

5.4.1 Stationary setting

We compute the number of appointment slots ($\mathbb{E}[G]$) that the oncologist must offer for a variety of arrival rates (λ) and waiting times ($\mathbb{E}[W]$). This data is displayed in Table 5.1 where cells contain the $\mathbb{E}[G]$ value corresponding to λ (row label) and $\mathbb{E}[W]$ (column label). It is clear from Table 5.1 that the number of appointments that need to be offered by the oncologist is very sensitive to the arrival rate of new patients. It is also apparent that offering a few extra appointment

λ	$\mathbb{E}[W] = 0.5$	$\mathbb{E}[W] = 1$	$\mathbb{E}[W] = 2$	$\mathbb{E}[W] = 6$
4	144.6	143.1	142.0	140.9
5	179.3	178.3	177.1	176.0
6	214.5	213.1	212.0	210.9
7	249.7	248.4	247.2	265.9
8	284.9	283.3	282.1	280.9
9	320.1	318.4	317.2	315.9

Table 5.1 – Appointment slots per month ($\mathbb{E}[G]$) for a given λ and $\mathbb{E}[W]$

slots per month can result is much improved waiting times.

The oncologist from which the data is based, saw on average 8.6 new patients per month and on average used 294.8 appointment slots per month (note that the panel size was not readily available). The average waiting time for patients was estimated to be between 0.5 and 2 months. In comparison, with $\lambda = 8.6$ and $\mathbb{E}[W] = 1$ as model inputs, the model computes the mean number of appointment slots per month to be $\mathbb{E}[G] = 304.6$ with utilization $\rho = 0.99$. It follows that the model predicts the oncologist will on average use $301.5 = 304.6 \cdot 0.99$ appointment slots per month. This represents an error of less than 5% when compared to observed data. This error is reasonable given one expects a certain degree of error in the observed data. For the remainder of the analysis in this section, model input parameters $\mathbb{E}[W] = 1$ and $\mathbb{E}[G] = 304.6$ are used.

At BCCA there are two operational policies being considered in an effort to allow an oncologist to see more new patients. The first policy change (Policy 1) is to have patients meet with nurse practitioners (instead of the oncologists) for some of the appointments. The second policy change (Policy 2) is to discharge patients more quickly (i.e. outsource some of the follow-up care to the patient's GP). Both of these policies are investigated with the described model.

We represent Policy 1 in our model by decreasing the frequency of follow-up appointments. Let q be this new frequency of follow-up

Policy 1	Policy 2: $\mathbb{E}[\hat{D}^i_{ap}] = \mathbb{E}[D^i_{ap}] - w$ $w = 0$	$w = 1$	$w = 2$	$w = 3$
$q = 1/i \cdot 1$	8.6	8.9	9.3	9.6
$q = 1/i \cdot 2/3$	12.9	13.4	13.9	14.5
$q = 1/i \cdot 1/2$	17.2	17.9	18.6	19.3

Table 5.2 – Mean arrival rates (λ) for new patients corresponding to changes to Policies 1 and 2

appointments, where $q = 1/i \cdot x$ and where x is the fraction of appointments handled by the oncologist. For example, when $x = 2/3$ then the nurse practitioner replaces the oncologist at every third appointment. We represent Policy 2 in our model by replacing $\mathbb{E}[D^i_{ap}]$ with $\mathbb{E}[\hat{D}^i_{ap}]$ where $\mathbb{E}[\hat{D}^i_{ap}] = \mathbb{E}[D^i_{ap}] - w$. Where w is the number of months sooner (on average) that a patient is discharged as compared to current practice.

To illustrate the impact of these policies, we consider values for parameters x and w to be $w = 0, 1, 2, 3$ and $x = 1, 2/3, 1/2$. In Table 5.2 the new patient arrival rates corresponding to these policy changes are displayed. The top-left cell represents the existing situation ($\lambda = 8.6$, $\mathbb{E}[W] = 1$ and $\mathbb{E}[G] = 304.6$) where neither policy is in use (i.e. $x = 1,\ w = 0$). In the remaining cells, λ is shown for the corresponding x and w values, while parameters $\mathbb{E}[W] = 1$ and $\mathbb{E}[G] = 304.6$ are kept constant.

As expected, Table 5.2 indicates that both policies will allow the oncologist to see more new patients. Policy 1, utilizing a nurse practitioner, results in the biggest gains allowing a significant increase in the arrival rate. Implementing Policy 2 will also allow an increase in the arrival rate. although to a lesser extent. Also evident from Table 5.2 is that when combined, Policies 1 and 2 can allow the arrival rate for new patients to increase by more than a factor of 2.

5.4.2 Non-stationary setting

In this subsection we compute the mean arrival rate for two non-stationary situations and plot the resulting demand for appointment slots and the resulting panel size. The first situation is when there is an unlimited number of new patients. This could represent, for example, the case where there is significant backlog at a clinic when the new oncologist starts. The second situation we consider is where there is limited number of new patients. By comparing the two situations we show the impact of a limited supply of patients. Finally to illustrate the relationship between the stationary and non-stationary, results from the stationary problem instance ($\lambda = 8.6$ and $\mathbb{E}[PS] = 2016$) are included in the relevant figures of this section.

Unlimited number of new patients: From (5.8) we can compute the maximum arrival rate for new patients that does not cause overtime frequency to be greater than φ. Applying (5.8) with $\varphi = 0.2$ and $g = 304.6$ leads the numeric results displayed in Figures 5.4, 5.5 and 5.6. Figure 5.4 plots the mean arrival rates for patients, Figure 5.5, the expected panel sizes and Figure 5.6, the total appointment slots and also the proportion of new and follow-up appointment slots.

From Figure 5.4 it is clear that a significant number of new patients can be accommodated by the new oncologist in the first months. However, that amount decreases rapidly during the first four years, after which, the oncologist can accommodate only a small amount more than in the stationary setting.

Figure 5.5 illustrates that the stationary panel size can be reach in approximately 144 months. This corresponds to the point in time when $\lambda(t) \approx 8.6$. Assuming the oncologist never accepts fewer than 8.6 new patients per month (λ from the stationary analysis), month 144 is the point in time when a backlog of patient begins to grow and when the stationary model becomes a better representation of the system.

From Figure 5.6 we see that the oncologist is always fully utilized.

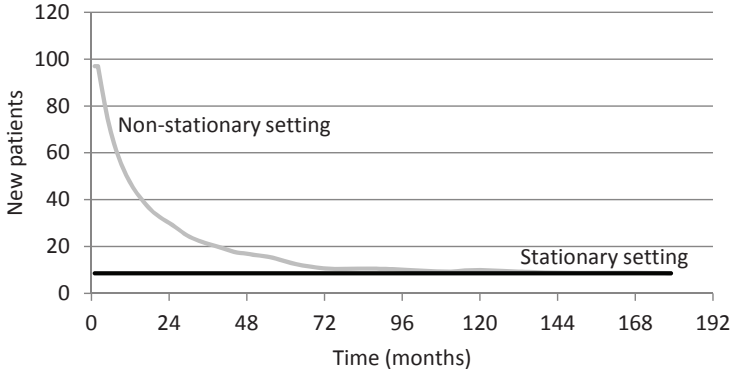


Figure 5.4 – Capacity to accept new patients (Unlimited number of new patients)

Fully utilized means that $\mathbb{E}[G(t)] = 291$ for all months, since when $\mathbb{E}[G(t)] > 291$ the frequency of overtime would exceed φ. This figure also clearly illustrates that in the first year the majority of the oncologist's time will be spent seeing new patients.

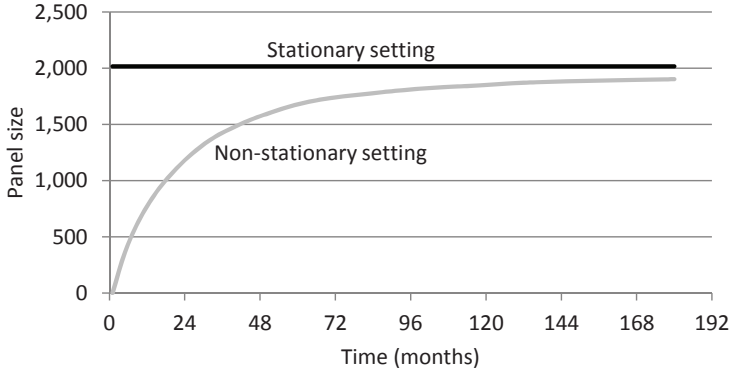


Figure 5.5 – Resulting panel size (Unlimited number of new patients)

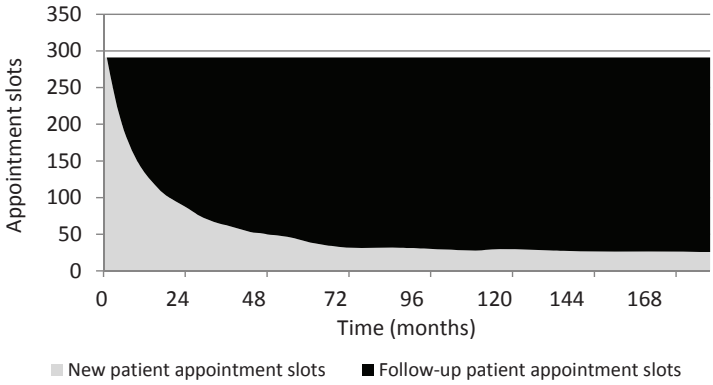


Figure 5.6 – Total appointment slots (Unlimited number of new patients)

These figures illustrate the value of both models for capacity planning. In particular, Figure 5.4 states the capacity of the oncologist to see new patients. Note however that this assumes an unlimited number of new patients and as such it is the maximum capacity that can be expected of a new oncologist.

Limited number of new patients: Incorporating a limited number of new patients can be accommodated by specifying a $\bar{\lambda}(t)$ value, see (5.8). To illustrate the impact, we set $\bar{\lambda}(t) = 50$.

From Figure 5.7 it is clear that during the first 2 years the oncologist will have capacity to see all 50 of the newly arriving patients. Afterward, that amount drops quickly, although it takes a significant amount of time until $\lambda(t) = \lambda = 8.6$. This is also clearly seen in Figure 5.8 where the slow panel size growth is observed. However, within two years the oncologist reaches maximum utilization as illustrated in Figure 5.9.

It follows, that even with a limited supply of new patients a new oncologist can be fully utilized quickly. Although, it take significantly longer for the oncologist to reach his/her stationary panel size. This

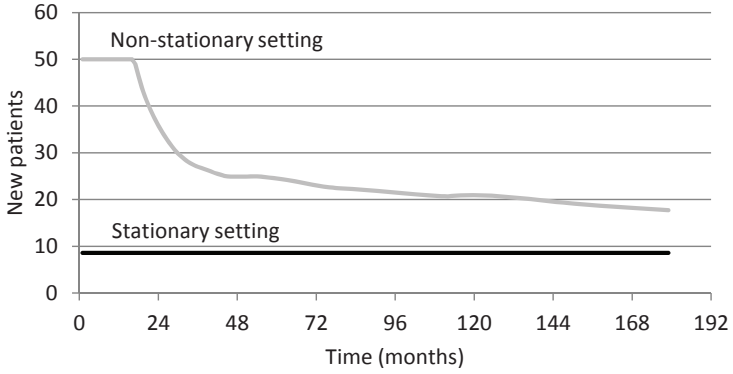


Figure 5.7 – Capacity to accept new patients (Limited number of new patients)

shows that it is advantageous for a new oncologist to see as many new patients as possible, as early as possible and thus the stationary panel size will be reached most quickly. Strategies to accomplish this may include 1) a transfer of patients (to the new oncologist) from established oncologists and/or 2) a period where all of the health centre's new patients go to the new oncologist and/or 3) a temporary increase in appointment slots during the early months of the new oncologist's clinic. Determining good strategies for incorporating the new oncologist into the health centre in a seamless manner, depends on the setting but is a natural extension of this research and a topic for further study.

5.5 Discussion

In this chapter we have formally defined how to extend existing panel size models to situations where the panel size in a random variable and where there are multiple patient and appointment types. In making these extensions, additional information was needed, namely the distribution of the panel size random variable. This distribution is specific to the application but as illustrated, it can be determined using

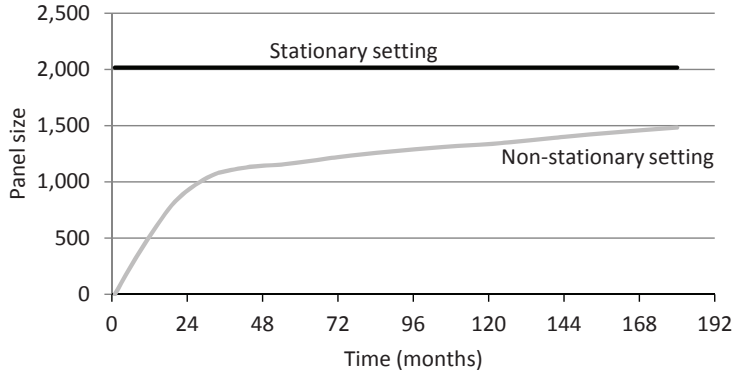


Figure 5.8 – Resulting panel size (Limited number of new patients)

queueing theory. For the oncology application we used two queueing network models representing a stationary and non-stationary system. The concept was illustrated to support capacity planning decisions at the BCCA. In addition to capacity planning, the models can be used in real-time to help an oncologist determine the number of new patients that should be seen.

Some further research could focus on other strategies to better utilize a new oncologist. For example, a different patient mix in the early years may lead to a populated clinic more quickly. Additionally it may be that oncologists change how they practice as they become more experienced. Including these changes when modelling may lead to further improvements.

Besides oncology, the methods described in this chapter can be used in other settings, particularly those treating chronic diseases. Depending on the setting and the performance metrics of interest, different queueing models (or simulation models) may be necessary to determine the panel size random variable. When the queueing model is tailored to the specific setting, the approached of this chapter can be used to support capacity planning, to determine operational confines of a clinic and to improve the startup period for a new physician.

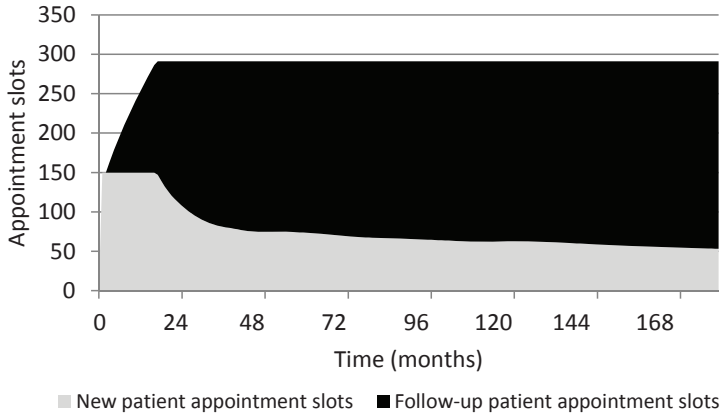


Figure 5.9 – Total appointment slots (Limited number of new patients)

5.6 Appendix

The technique used to aggregate the parameters in the multi-class opening queueing network is described in this appendix. We only include formulations specific for the model of Subsection 5.3.1. See [17] for more details on the method.

Let $J = \{np, ap, ip\}$ be the set of queues in the network and let $j \in J$. We start by computing the first and second moment of the aggregate service time (D^+_{ap}) for patients in the active patient queue. This amounts to a weighted average of the service time of each patient type and is computed as follows,

$$\begin{aligned} \mathbb{E}[D^+_{ap}] &= \frac{1}{\sum_{i=1}^{c} \lambda^i_{ap}} \sum_{i=1}^{c} \lambda^i_{ap} \mathbb{E}[D^i_{ap}] \\ VAR[(D^+_{ap})] &= \frac{1}{\sum_{i=1}^{c} \lambda^i_{ap}} \sum_{i=1}^{c} \lambda^i_{ap} VAR[(D^i_{ap})] \end{aligned}$$

where $VAR[x]$ is the variance of random variable x and λ^i_j is the

arrival rate of patient type i to queue j defined as $\lambda^i_{np} = \lambda p_i$, $\lambda^i_{ap} = \lambda^i_{np} + \lambda^i_{ip} r^i_{ip,ap}$ and $\lambda^i_{ip} = \lambda^i_{ap} r^i_{ap,ip}$.

From these aggregate values the squared coefficient of variance for the service time in the active patient queue ($SCV_{ap,s}$, note that subscript ap indicates the active patient queue and subscript s indicates service) can be obtained as follows,

$$SCV_{ap,s} = \frac{1}{\lambda^+_{ap}(\mathbb{E}[D^+_{ap}])^2} \times \sum_{i=1}^{c} \left(\lambda^i_{ap} \mathbb{E}[D^i_{ap}]^2 \left(\left(\frac{\sqrt{VAR[D^i_{ap}]}}{(\mathbb{E}[D^i_{ap}])} \right)^2 + 1 \right) - 1 \right). \quad (5.9)$$

The aggregate mean arrival rate and the aggregate routing probabilities are respectively, $\lambda^+_j = \sum_{i=1}^{c} \lambda^i_j$ and $r^+_{k,j} = 1/\lambda^+_k \sum_{i=1}^{c} \lambda^i_k r^i_{k,j}$, $k \in J$. Since all new patients go to queue np, the external (new) patient arrival rate $\lambda^+_0 = \lambda$ and $r^+_{0,j} = 1$, when $j = np$ and 0 otherwise.

At this point the c patient classes are aggregated into a single class and we now consider the network to be a single class open queueing network with the aggregate parameters described above. To analyze the single class open queueing network we next determine the SCV for the arrival processes to each queue ($SCV_{j,a}$) as follows,

$$\begin{aligned} SCV_{np,a} &= \alpha_{np} \\ SCV_{ap,a} &= \alpha_{ap} + SCV_{np,a}\beta_{np,ap} + SCV_{ip,a}\beta_{ip,ap} \\ SCV_{ip,a} &= \alpha_{ip} + SCV_{ap,a}\beta_{ap,ip} \end{aligned} \quad (5.10)$$

where α_j and $\beta_{i,j}$ are constants depending on the input data,

$$\begin{aligned}\alpha_j = \quad & 1 + w_j \left(\pi_j \left(\frac{\sqrt{VAR[N]}}{\lambda} \right)^2 - 1 \right) + \\ & w_j \left(\sum_{k \in J} \frac{\lambda_k^+ r_{k,j}^+}{\lambda_j^+} \left((1 - r_{k,j}^+) + r_{k,j}^+ \rho_k^2 x_k \right) \right) \\ \beta_{k,j} = \quad & w_j r_{k,j}^+ \frac{\lambda_k^+ r_{k,j}^+}{\lambda_j^+} (1 - \rho_k^2), \ k \in J\end{aligned}$$

and $\pi_{np} = 1, \pi_{ap} = \pi_{ip} = 0$ and

$$\begin{aligned} w_j &= \left((1 + 4(1 - \rho_j)^2 (v_j - 1) \right)^{-1}, \\ v_j &= \left(\sum_{k \in J \cup \{0\}} \left(\frac{\lambda_k^+ r_{k,j}^+}{\lambda_j^+} \right)^2 \right)^{-1}, \quad \rho_j = \frac{\lambda_j^+ \mathbb{E}[D_j^+]}{m_j} \end{aligned}$$

and

$$x_j = 1 + m_j^{-0.5} (\max[SCV_{j,s}, 0.2] - 1).$$

Since m_{np} and m_{ip} are large, then $x_{np} = x_{ip} \approx 1$ and $\rho_{np} = \rho_{ip} \approx 0$

Using the SCV for the arrival process (5.10) and the SCV for the service time (5.9), the expected waiting time in the active patient queue is approximated as follows (see [169]),

$$\mathbb{E}[W] \approx \frac{SCV_{ap,a} + SCV_{ap,s}}{2} \frac{\rho_{ap}^{\left(\sqrt{2(m_{ap}+1)}-1\right)}}{m_{ap}(1 - \rho_{ap})} \mathbb{E}[D_{ap}^+].$$

Chapter 6

Surgical scheduling and inpatient wards

Contents

6.1 Introduction

No other single hospital department influences the workload of other departments more than the Department of Surgery, and in particular, the activities in the operating room (OR) [127]. This influence depends directly on what types of patients receive surgery and when. Generally speaking, more invasive surgeries require more care during a patient's recovery. Giving consideration to the downstream effect of the OR is essential for balancing the workload of the hospital. The planning and

scheduling of OR time is discussed by many authors [14, 20, 22, 35, 96, 197] and is often described as a multiple stage process.

The multiple stage process used by many hospitals starts with the long term allocation of OR time to surgical specialties, e.g. the number of surgery hours per year. This allocation, referred to as Stage 1, is a strategic decision that reflects patient demand patterns and the priorities defined by hospital management. From this strategic decision a master surgical schedule (MSS) is developed for a shorter time horizon which divides OR time (aggregated into blocks) amongst the specialties; known as Stage 2. The specific assignment of patients to OR blocks within the MSS is commonly referred to as Stage 3. A fourth stage "addresses the monitoring and control of the OR activities" [157] on the day of surgery. In this chapter we focus on the development of a MSS in Stage 2.

The MSS is often specialty specific [14] meaning OR time is dedicated to a surgical specialty. In these MSSs, the decision of which patients (and consequently which surgeries) to schedule within each OR block, is determined by the surgical specialty through consultation with the OR manager. Other MSSs are more specific with OR blocks being allotted to specific surgical procedures [97, 158]. Instead of using the term MSS, other authors refer to the distribution of OR time amongst surgical specialties as a surgical block schedule [172] and a timetable of operations [92].

The development of a MSS is often a complex balancing act. Since the OR is one of the hospital's most expensive resources, hospitals wish to maximize its performance through high resource utilization, minimal overtime, minimal cancellations and the elimination of conflicting equipment needs between rooms. Many authors describe methods for developing the MSS that take into account various resources within the OR such as staff, equipment and instrument trays. For a review see [35]. Furthermore, the OR is often described as the engine that drives the hospital [127], implying many other departments are impacted by the MSS. The effect of the MSS on ward occupancy [1, 14–16, 78, 79, 92, 97, 100, 158, 187–189], critical care re-

sources [15, 78, 82, 97, 142, 158] and waiting lists [172, 189] has notably been studied. Three of the mentioned papers represent the relationship with deterministic models, while the remaining consider at least one variable as stochastic. The stochastic models are either simulation models, mathematical programming models or queueing theory models. In the following paragraph we relate the model presented in this chapter to these techniques, thereby highlighting the contribution of our approach to MSS development.

Simulation models [82, 92], which are well equipped to capture the broad scope of complex systems, typically require a great deal of time to develop. As such, when analytical approaches can be used as an alternative, they have distinct advantages in terms of development time. Mathematical programming models [1, 172] on the other hand are capable of optimizing for certain objectives but require a more limited scope in order to be solvable. Queueing theory models [78, 79, 100, 188] are particularly adept at coping with system variability, however typically also employ a more limited scope than simulation. As such, many of the factors that may be included in simulation models are ignored in mathematical programming and queueing theory models.

The analytical model presented in this chapter, which most closely resembles a queueing model, has advantages in terms of development time. Furthermore, to broaden the scope of mathematical programming models, our model can be used to quickly evaluate proposed MSS solutions for a number of additional factors. Such evaluations can be automated and a feedback loop can be created allowing the original mathematical programming model to account for additional factors.

Using our model, downstream workload distributions can be computed as a function of the MSS for all departments that provide care for recovering surgical patients. Specifically the model computes the ward occupancy distributions, the patient admission / discharge distributions and the distributions for the ongoing interventions / treatments required by recovering patients. Furthermore, the cumulative influence of multiple MSS cycles are considered. Since the MSS is identical from cycle to cycle, the overlapping of patients from one cycle to the

next can be anticipated. A single MSS design is expected to remain in place for a long period of time leading to "steady-state" workload distributions for each day of the MSS cycle. From a tactical point of view, this allows managers to evaluate the workload for a given MSS. Alternatively, the same model can support decisions at an operational level. Instead of computing the expected patients in recovery, the actual patients in recovery can be used as input. By agreggating this with the expected new arrivals from the OR, real-time workload projections can be used to identify upcoming staffing needs.

Such a model is valuable for department managers who want a means to relate their department's workload to the activities of the Department of Surgery. The motivation for this model comes from a real case at NCI. NCI planned to open an additional OR and wanted to know the impact of this decision on the hospital as a whole, not only on the Department of Surgery. The model presented in this chapter was applied at NCI to support the development of their new MSS.

The chapter is organized as follows. Initially we describe the model and how to derive workload metrics. The next section describes the application of the model and the results at NCI. Section 6.4 discusses the use of the model as part of a decision support tool. The final section offers concluding remarks.

6.2 Model description

This section describes a model to determine the workload placed on hospital departments by recovering surgical patients. In the same way a MSS describes resource demands within the OR, we show how the resources of other departments can be seen as a function of the MSS. The method relies only on data which are easily extractable from typical patient management systems.

The model is most easily described from a queueing theory perspective. The basic component of the model is a single OR block and its expected impact on the arrival rate to the hospital wards. The number of cases

scheduled in such a block varies per specialty and is modelled as a specialty specific random variable. This variable also represents the number of patients arriving to the ward (batch size). At the ward, each patient directly occupies a bed for a certain period of time. In the queueing model, the ward is seen as an infinite server system in which the patients occupy a server (ward bed) without delay. The time spent occupying a bed (length of stay (LOS)) is the service time, which is modelled as a random variable. Again, this random variable is specific to the surgical specialty. Since patients occupying a server do not interfere with each other during their recovery, the aggregate number of patients for all OR blocks can be computed by adding the individual effects of all OR blocks. Finally, since the MSS is cyclical, the cumulative number of patients from recurring MSS cycles can be computed.

The main output of the model is the distribution for the number of patients anticipated in the system on each day of the MSS. The model used for these calculations is explained in the following subsection. The three subsequent subsections explain how the model can be modified to obtain the distributions for 1) ward occupancies, 2) admissions/discharges and 3) the number of patients in a specific day of their recovery. The time scale in the model is days; therefore all metrics are considered on a daily basis.

6.2.1 Model inputs

A MSS represents a repetitive pattern over a certain number of days (say Q). For each day $q \in \{1, 2, ..., Q\}$ in the MSS each of the I available ORs can be assigned to one of the available surgical specialties. More precisely, the MSS is described by the assignment of a surgical specialty j to each OR block $b_{i,q}$ where $i \in \{1, 2, ..., I\}$. Using this notation, an empty MSS (i.e. before specialties have been assigned to OR blocks) is shown in Figure 6.1 where each cell represents an OR block. It is possible for multiple blocks to be assigned to a single specialty on the same day.

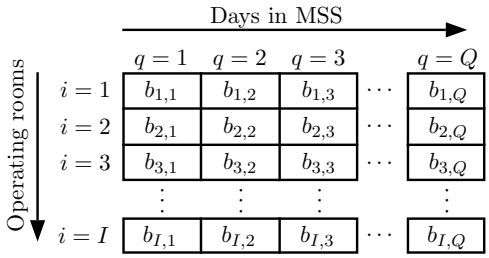


Figure 6.1 – Empty MSS illustrating chapter notation

The way specialty j fills in an OR block is described by two parameters, c^j and d_n^j. Parameter c^j is a discrete distribution for the number of surgeries carried out in one block, i.e. $\mathbb{P}(c^j = k)$ is the probability of k surgeries, $k \in \{0, 1, ..., C^j\}$, where C^j is the maximum number of surgeries of specialty j that can be completed in one block. Specialties independently decide which patients to schedule during each block, meaning that the number of surgeries completed in one block does not influence the number of surgeries completed in another. The second parameter d_n^j is the probability that a patient, who is still in the ward on day n, is to be discharged that day ($n \in \{0, 1, ..., L^j\}$, where L^j is the maximum LOS for specialty j, a finite LOS is used for numerical purposes). Note that d_0^j is the probability that the patient is discharged on the same day as surgery (i.e. an outpatient surgery or day-case surgery) and $d_{L^j}^j = 1$. The parameter d_n^j is computed by dividing the probability that a patient's total stay is exactly n days by the probability that the patient was not yet discharged before day n. Let $P^j(n)$ be the probability that the LOS of a patient from specialty j is exactly n days long, then formally d_n^j is computed as follows,

$$d_n^j = \frac{P^j(n)}{\sum_{k=n}^{L^j} P^j(k)}. \tag{6.1}$$

6.2.2 Recovering patients in the hospital

Using c^j and d_n^j as model inputs, for a given MSS the probability distribution for the number of recovering patients on each day q is computed with three steps. Step 1 computes the distribution of recovering patients from a single OR block of a specialty j, i.e. we essentially precalculate the distribution of recovering patients expected from an OR block of a specialty. In Step 2, we consider a given MSS and use the result from Step 1 to compute the distribution of recovering patients given a single cycle of the MSS. Finally in Step 3 we incorporate recurring MSSs and compute the probability distribution of recovering patients on each day of the MSS.

Step 1: Distribution of recovering patients from specialty j following from a single OR block: In Step 1 we ignore the MSS and consider a single specialty j operating in a single OR block. The patient flow process is as follows. During the OR block patients receive surgery. The number of patients who undergo surgery in one OR block is given by the random variable c^j. After surgery each patient still on the ward on day n has the probability d_n^j of being discharged that day. In the following we compute the probability $\mathbb{P}(h_n^j = x)$ that n days after carrying out a block of specialty j, x patients of the block are still in recovery. Note that $n \in \{0, 1, ..., L^j\}$ and $x \in \{0, 1, ..., C^j\}$ and that, for example, $\mathbb{P}(h_3^j = 5) = 0.25$ means that 3 days after surgery there is a 25% chance that 5 patients are still recovering in the hospital.

Day $n = 0$ is defined as the day of surgery and it is assumed that patients occupy a bed all day on the day of surgery even though they may physically be in the OR. This is consistent with practice where patients have a recovery bed reserved for them before surgery. As such the number of patients in recovery from specialty j on day $n = 0$ is by definition the number of surgeries performed that day by specialty j. It follows that the distribution for the number of recovering patients on day $n = 0$ is $h_0^j = c^j$.

Note that on day n, each patient still in the hospital has a probability d_n^j of being discharged that day and $(1 - d_n^j)$ of staying. If there are

k patients in recovery on day n, then the probability of s patients in recovery (where $s \leq k$) on day $n+1$ is computed using the binomial distribution, $\binom{k}{s}(d_n^j)^{k-s}(1-d_n^j)^s$. Since we know the probability distribution for the number of patients at the end of day $n=0$ we can iteratively use this formula to compute the probability of k patients at the end of all days $n > 0$. Summarizing, the distribution for the number of recovering patients on day n is recursively computed by,

$$\mathbb{P}(h_n^j = x) = \tag{6.2}$$
$$\begin{cases} \mathbb{P}(c^j = x) & \text{when } n = 0 \\ \sum_{k=x}^{C^j} \binom{k}{x} (d_{n-1}^j)^{k-x}(1-d_{n-1}^j)^x \mathbb{P}(h_{n-1}^j = k) & \text{otherwise.} \end{cases}$$

Step 2: Aggregate distribution of recovering patients following from a single MSS cycle: In this step we consider the previously computed probability distribution h_n^j and a given MSS as input. Although the MSS is cyclical and repeats after Q days, in this subsection we consider only a single MSS cycle in isolation. The MSS defines when each specialty is assigned an OR block and thus the days on which patients of specialty j arrive to the ward. Based on these, we compute the total number of patients in recovery by means of discrete convolutions.

To calculate the overall distribution of recovering patients, we first identify for each block $b_{i,q}$ the impact that this block has on the number of recovering patients in the hospital on days $(q, q+1, ...)$. If z denotes the specialty assigned to block $b_{i,q}$ which follows from the MSS, then the distribution $\bar{h}_m^{i,q}$ for the number of recovering patients of block $b_{i,q}$ on day m ($m \in \{1, 2, ..., Q, Q+1, Q+2, ...\}$) is given by,

$$\bar{h}_m^{i,q} = \begin{cases} h_{m-q}^z & \text{if } q \leq m < L^z + q, \\ \mathbf{0} & \text{otherwise} \end{cases} \tag{6.3}$$

where $\mathbf{0}$ means $\mathbb{P}(\bar{h}_m^{i,q} > 0) = 0$. Note that specialties index j is no longer needed as specialties are accounted for by their designated OR block(s).

Let H_m be a discrete distribution for the total number of recovering patients on day m resulting from a single MSS cycle. Since recovering patients do not interfere with each other we can simply iteratively add the distributions of all the blocks impacting day m to get H_m. Adding two independent discrete distributions is done by discrete convolutions which we indicate by $*$. Let A and B be two independent discrete distributions. Then $C = A * B$, is computed by,

$$\mathbb{P}(C = x) = \sum_{k=0}^{\tau} \mathbb{P}(A = k)\mathbb{P}(B = x - k)$$

where τ is equal to the largest x value with a positive probability that can result from $A * B$. Using this notation, H_m is computed by,

$$H_m = \bar{h}_m^{1,1} * \bar{h}_m^{1,2} * ... * \bar{h}_m^{1,Q} * \bar{h}_m^{2,1} * ... * \bar{h}_m^{I,Q}. \tag{6.4}$$

Step 3: Steady-state distribution of recovering patients: In Step 3 we consider a series of MSSs to compute the steady-state probability distribution of recovering patients. The cyclic structure of the MSS implies that patients receiving surgery during one cycle may overlap with patients from the next cycle. In the case of a small Q for example, patients from many different cycles can overlap.

In Step 2 we have computed H_m for a single cycle of the MSS in isolation. Let M be the last day where there is still a positive probability that a recovering patient is present as computed by H_m. Thus $M = \max_j\{L^j + x^j\}$ (where x^j is the latest day q of a block assigned to specialty j) indicates the range of the MSS. To calculate the overall distribution of recovering patients when the MSS is repeatedly executed we must take into account $\lceil M/Q \rceil$ consecutive cycles of the MSS (see Figure 6.2). Let H_q^{SS} denote the probability distribution of recovering patients on day q of the MSS cycle, resulting from $\lceil M/Q \rceil$ consecutive MSS cycles. Since the MSS does not change from cycle to cycle, H_q^{SS} is the same for all MSS cycles. Using discrete convolutions, H_q^{SS} is computed by,

$$H_q^{SS} = H_q * H_{q+Q} * H_{q+2Q} * ... * H_{q+\lceil M/Q \rceil Q}. \tag{6.5}$$

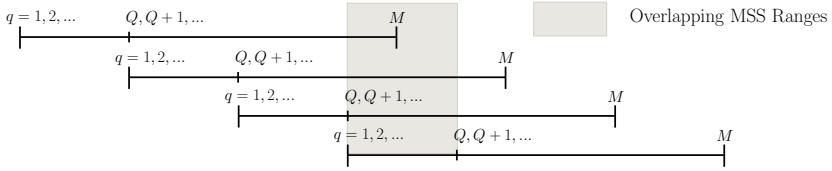


Figure 6.2 – Consecutive MSS cycles illustrating overlapping recovering patients

The relationship between the distribution H_q^{SS} and the work load associated with recovering patients is discussed in detail in the following three subsections.

6.2.3 Ward occupancy

Perhaps the most common measure of inpatient workload is ward occupancy. Ward occupancy, the distribution of the number of inpatients on a ward, follows easily from the basic model where we compute the distribution of all recovering patients. In practice patients tend to be segregated into different wards depending on the type of surgery they receive. To incorporate this segregation into the model and to consequently have recovering patient distributions for each ward, a minor modification needs to be made to the model. Let W_k be the set of specialties j whose patients are admitted to Ward k. Then in Step 2 we only have to consider those OR blocks assigned to a specialty in W_k.

6.2.4 Admission rate / discharge rate

Bed occupancy alone does not fully account for the workload associated with caring for recovering patients. During patient admissions and discharges the nursing workload can increase. As such in this subsection we explain how to derive the probability distribution for daily admissions and discharges.

The admission rate is the rate of arriving patients which we previously defined as the number of surgeries completed on day $n = 0$. For this metric we are only interested in a patient on the day of admission and wish to ignore them afterward. To modify the model to reflect this new purpose replace (6.2) with,

$$h_n^j = \begin{cases} c^j & \text{when } n = 0 \\ \mathbf{0} & \text{otherwise.} \end{cases} \tag{6.6}$$

With this modification, the resulting H_m represents the distribution for daily admission for each day q of the MSS. To have ward specific results we again can restrict this to blocks belonging to specialties of the specific ward.

The discharge rate is the rate at which patients leave the ward and can be computed by adding an additional calculation in Step 1. The number of patients in recovery on day n is distributed according to h_n^j, see (6.2). On day n each patient has the probability d_n^j of being discharged and the probability $(1-d_n^j)$ of staying. Let D_n^j be a discrete distribution for the number of discharges from specialty j on day n. Given h_n^j and d_n^j, D_n^j can be computed with a binomial distribution as follows,

$$\mathbb{P}(D_n^j = x) \quad = \quad \sum_{k=x}^{C^j} \binom{k}{x} (d_n^j)^x (1 - d_n^j)^{k-x} \mathbb{P}(h_n^j = k). \tag{6.7}$$

Finally, after computing D_n^j, one can set $h_n^j = D_n^j$ and continue with Step 2. By doing so, the resulting H_q^{SS} represents the distribution for daily discharges for each day q of the MSS. As with admissions, ward specific results can also be obtained.

6.2.5 Patients in day n of their recovery

The final workload metric we consider is the distribution of patients in day n of their recovery. This is relevant for predicting workload for the many hospital departments who treat recovering patients. For

example, a patient recovering from hip surgery may need to see a physiotherapist every other day during their recovery. This metric states the frequency of such visits. The analogy holds true for all types of services that take place on specific intervals during a patient's recovery (e.g. chemotherapy, diagnostics, social work, discharge planning). In particular, this metric can help dimension capacity for clinical pathways patients whose recovery is intended to follow a strict regime.

The metric requires substantial modifications to the original model, since we now must carry an index (n) for the 'day of recovery' throughout the three steps. Let $\bar{h}^{i,q}_{m,n}$ be a discrete distribution for the number of recovering patients from block $b_{i,q}$ on day m in day n of their recovery. To compute $\bar{h}^{i,q}_{m,n}$ replace (6.3) with the following,

$$\bar{h}^{i,q}_{m,n} = \begin{cases} \mathbf{0} & \text{if } m - q \neq n \\ h^{z}_{m-q} & \text{otherwise,} \end{cases} \tag{6.8}$$

and we replace (6.4) with the following,

$$H_{m,n} = \bar{h}^{1,1}_{m,n} * \bar{h}^{1,2}_{m,n} * ... * \bar{h}^{1,Q}_{m,n} * \bar{h}^{2,1}_{m,n} * ... * \bar{h}^{I,Q}_{m,n} \tag{6.9}$$

where $H_{m,n}$ now denotes the number of patients from a single MSS on day m in day n of their recovery.

This alteration to the model eliminates the need for convolutions in Step 3. Since patients are indexed by their recovery day, patients from one MSS cycle are not aggregated with patients from the next. As such we need to replace (6.5) with,

$$H^{SS}_{q,n} = H_{q+Q\lfloor n/(Q+1)\rfloor,n}. \tag{6.10}$$

To help to interpret this metric, consider the following fictitious example for patients who require chemotherapy treatment on day 3 of their recovery. The Chemotherapy Department would like to know how frequently they need to provide this service. Example results for $H^{SS}_{q,n}$ are illustrated in Table 6.1.

Example results		Interpretation
$\mathbb{P}(H_{1,3}^{SS} = 2) = 0.3$	$n = 3,\ q = 1$	30% probability that exactly 2 treatments are required on the first day of the MSS cycle
$\mathbb{P}(H_{1,3}^{SS} = 3) = 0.5$	$n = 3,\ q = 1$	50% probability that exactly 3 treatments are required on the first day of the MSS cycle
$\mathbb{P}(H_{1,3}^{SS} = 4) = 0.2$	$n = 3,\ q = 1$	20% probability that exactly 4 treatments are required on the first day of the MSS cycle
$\mathbb{P}(H_{2,3}^{SS} = 2) = 0.4$	$n = 3,\ q = 2$	40% probability that exactly 2 treatments are required on the second day of the MSS cycle
$\mathbb{P}(H_{2,3}^{SS} = 3) = 0.4$	$n = 3,\ q = 2$	40% probability that exactly 3 treatments are required on the second day of the MSS cycle
$\mathbb{P}(H_{2,3}^{SS} = 4) = 0.2$	$n = 3,\ q = 2$	20% probability that exactly 4 treatments are required on the second day of the MSS cycle

Table 6.1 – Example results for the frequency of inpatient chemotherapy treatments

6.2.6 Assumptions

Inherent to the model are a number of assumptions which are discussed in this subsection. One assumption resulting from the use of the infinite server system, is that there is always a bed available for a patient after surgery. This implies that surgeries are never cancelled due to bed shortages. In practice this means that there is not a physical bed restriction and that additional staff can be called in when demand is higher than expected. The frequency of this occurring follows from the model. For example if a hospital staffs 50 beds, then the probability of an additional staffed bed being needed on day q is $\mathbb{P}(H_q^{SS} = 51)$.

In the current formulation the model ignores seasonality. Of course at certain times of the year surgical blocks are cancelled to accommodate

vacations and slowdowns, representing a change in supply. In this case a modified MSS is temporarily used breaking down the assumption that the same MSS repeats every Q days. However, given that the modifications to the MSS are typically cancellations of certain OR blocks, then the original result can act as an upper bound.

Finally only elective surgeries are considered. To incorporate non-elective surgeries, it is possible to convolute a historic bed occupancy distribution for non-elective patients. Alternatively, it is possible to incorporate a virtual OR block into the model that represents emergency admissions.

The model output reflects the variability in both c^j and d^j. Higher variability in these parameters leads to higher variability in the results. As such the planning and scheduling becomes more difficult. In the NCI case study that follows, for most OR blocks the choice of which procedure to be carried out within the OR block was left to the specialty. In hospitals where OR blocks have fixed procedure assignments [97, 158], one expects less variability in c^j and d^j and thus less variability in the model output.

The inherent assumption of using the binomial distribution in this model is that all patients (experiments) have equal probability of each outcome and that the outcome is independent of other patients, i.e. it is assumed that the patients are independent and identically distributed (IID). The independence assumption is natural as it implies that the amount of time one patient is in the hospital does not influence the amount of time another patient is in the hospital. The identically distributed requirement means that we must compute the number of beds needed tomorrow (and the number of case completed in one OR block), for all identically distributed cohorts of patients separately. In other words, the parameters of the binomial distribution must reflect all of the patients in a given cohort (For a discussion on defining statistically equivalent patient cohorts, see [89]). In our model we aggregate patient such that each surgical specialty is a patient cohort. It follows then, that patients within each surgical specialty should be identically distributed.

If a heterogeneous population is grouped together, this causes the ward census distribution to have longer tails (although the mean remains the same) and will overestimate the bed requirements when staffing for a certain percentile of demand. On the other hand however, less aggregation (such as dividing a specialty by short and long stay patients) decreases the sample size from which to derive the parameters which in turn reduces the statistical confidence of the estimated parameters. In our case study that follows, we aggregate the data by specialty which allows for enough data to have a sufficient sample size and results in relatively homogeneous patient cohorts.

In cases where patients of a surgical specialty are not identically distributed and cannot be aggregated into a single cohort the model can still be used. First the heterogeneous specialty has to be divided into multiple homogeneous cohorts and then these cohorts can be treated as if they were assigned their own OR block. Using this, the binomial distribution is applied as described above to determine the bed requirements of each cohort. Again using the independence assumptions, these cohorts can be added (with discrete convolutions) to determine the total bed requirements for the complete surgical specialty.

6.3 Application

As with many Dutch hospitals, NCI is eager to improve access, and so has expanded its surgical capacity by one OR. The Department of Surgery was empowered to develop a new MSS encompassing the additional OR which fit within the capacity constraints of their department. The downstream wards, were asked to increase their staffing to accommodate the influx of extra patients. However, the extent to which they had to increase staff, and when they had to increase (in terms of which weekdays), was not known. In this section, we describe how the model was used to project the ward occupancy and to determine a new MSS for the hospital.

This section is divided into two subsections. The first subsection dis-

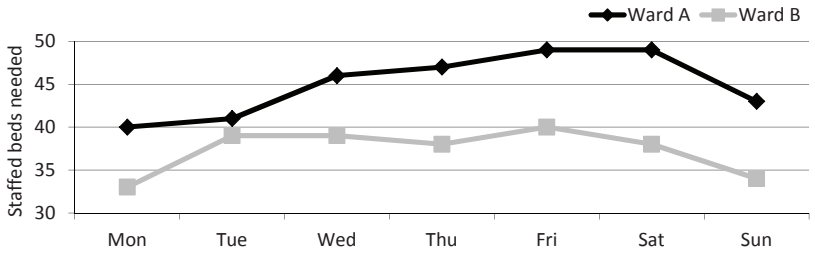


Figure 6.3 – 90^{th} percentile of demand projected by the model for each day of the MSS cycle (Original MSS proposal)

cusses the ward occupancy projected by the model during the MSS development process. We show ward occupancy projections from the original MSS proposal and from the MSS proposal that staff chose to implement (which we refer to as the implemented MSS). In the second subsection we compare the ward occupancy projected by the model for the implemented MSS with the ward occupancy observed after it was implemented.

6.3.1 Projected results

NCI has two wards for treating surgical patients, Ward A and Ward B, with a combined physical capacity of 100 beds. Management strives to staff enough beds such that for 90% of the days there is sufficient coverage. In other words, they staff for the 90^{th} percentile of demand; their accepted risk for needing to call in additional staff is thus 10%. Figure 6.3 illustrates the 90^{th} percentile demand for staffed beds on each of the wards, resulting from the original MSS proposal. As is clear from the figure, the staffing requirements are relatively balanced across the weekdays (Monday to Friday) for Ward B. This is not the case for Ward A. On Ward A the occupancy is relatively low on Monday and Tuesday, and relatively high on Thursday, Friday and Saturday.

This projected demand for staffed beds concerned the ward manager, as such an unbalanced demand profile makes staff scheduling, and ward

operations, difficult. Early in the week, beds would be underutilized whereas later in the week, beds would become highly utilized leading to significant problems, particularly as the wards approach peak capacity. For example, when inpatient wards reach their peak capacity and a patient admission is pending, staff often scramble to try and make a bed available. If one cannot be made available, additional staff are called in (or in rare cases when additional staff cannot be found, the elective surgery is cancelled), which causes extra work for OR planners, wasted time for surgeons and anxiety for patients. When a bed is made available, it often means a patient was transferred from one ward to another (often to a ward capable of caring for the patient but not the designated one) or discharged. Either way, extra work is required by ward staff and there is a disruption in patient care. Although completely eliminating such problems is likely not possible without an exorbitant amount of resources (due to the variance), sound planning ahead of time may help to minimize occurrences.

After discussing the model output, all participating staff agreed that the original MSS, although appropriate for the OR, was not ideal for the wards. The discussion then moved to how to correct the imbalance across the weekdays by changing the assignment of OR blocks to specialties. Modifications to the original MSS were made by considering what changes were possible within the restrictions of the OR (e.g. physician schedules and equipment availability).

Eventually, after considering several MSS proposals, the process led to a MSS (the implemented MSS) which was acceptable to all staff members. The implemented MSS fit within the restrictions of the OR and, as illustrated in Figure 6.4, resulted in a more balanced ward occupancy. Comparing the implemented MSS with the original MSS, the implemented MSS dampened the fluctuation on Ward A by lowering occupancy on Thursday, Friday and Saturday, and increasing it on Monday and Tuesday. With the implemented MSS the model predicted that no days would require more than 47 staffed beds, which reduced the maximum from 49 (predicted for the original MSS). Furthermore, the implemented MSS ensured the staffing requirements re-

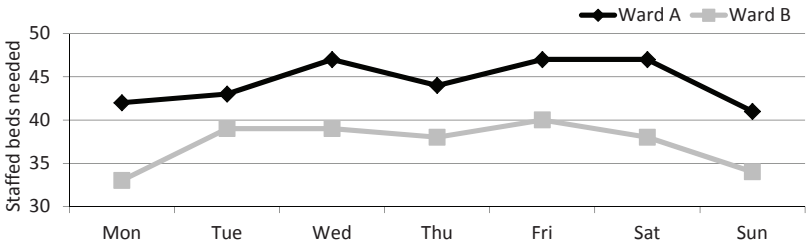


Figure 6.4 – 90^{th} percentile of demand projected by the model for each day of the MSS cycle (Implemented MSS)

mained relatively balanced across the working days for both wards.

6.3.2 Observed results

The ward occupancy was observed over a 33 week period after the new OR was fully operational. From these data, probability distributions of beds used for each day of the MSS cycle were derived. Using Chi-square goodness-of-fit tests [139], these observed distributions were compared to those projected by the model. For Ward B, six of the seven distributions (one for each day of the MSS cycle) were found to be statistically equivalent at a level $\alpha = 0.05$, while the seventh day was statistically equivalent at a level $\alpha = 0.2$. For Ward A, the tests revealed statistical equivalence at levels $\alpha = 0.15$ (for three days), 0.25 (for two days) and 0.35 (for two days). At these alpha levels we conclude that the observed ward occupancy is well predicted by the model. Explanations for the poorer fit of Ward A data is discussed in the following paragraphs where the 90^{th} percentiles (desired staffing level) are compared for the observed and projected results.

Figure 6.5 and Figure 6.6 compare the projected ward occupancy with the observed ward occupancy during the 33 week period. Figure 6.5 displays results for Ward A and Figure 6.6 for Ward B.

As is observable in Figure 6.5 and Figure 6.6 the data indicates that both wards have balanced ward occupancies across the week days.

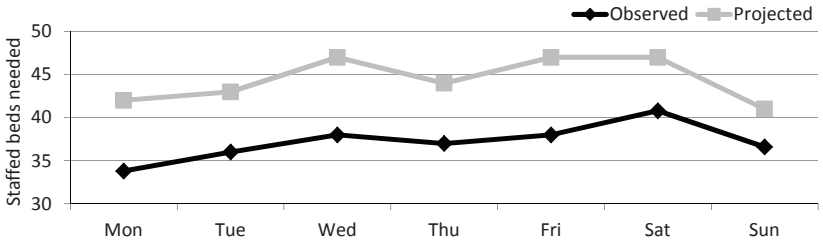


Figure 6.5 – Comparison of projected and observed ward occupancies (90^{th} percentile) on Ward A

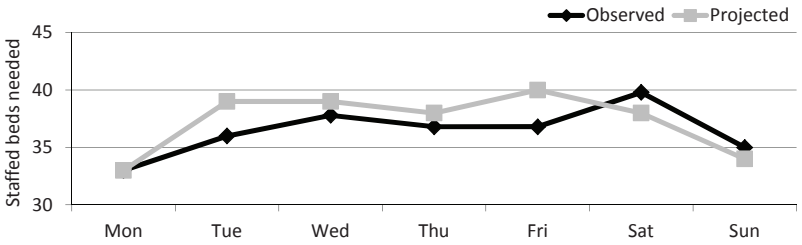


Figure 6.6 – Comparison of projected and observed ward occupancies (90^{th} percentile) on Ward B

However, it is also observable that our model overestimated the number of beds required in Ward A by approximately 16%. The overestimate is due to an unexpected increase in short stay patients during the period of measurement. Had this change in patient mix been expected at the time the projections were made (and model input altered to reflect it), such an overestimate would likely not have been observed, and we would expect to have similarly accurate results as those for Ward B. This highlights how process variables may change over time and the importance of estimating this and reflecting it in the model input.

As a final note on the model results, consider if hospital management decided to staff only for the average number of beds projected to be needed for 6 ORs. In this case, 32 beds would be assigned to Ward A

and 29 beds to Ward B. This would have led to a bed shortage on 51% of the days. This illustrates the importance of considering probability distributions in hospital planning.

6.4 Commercial software

In this subsection we discuss a commercial software application which has been developed based on the model described in this chapter. Although the software development is not a contribution of this thesis, the software does show that the research is not only of theoretical value, but also of practical value. Furthermore, this software is marketed beyond specialty cancer hospitals demonstrating that the approach is valid in other hospital settings.

To support future redesign of the MSS, the model described in this paper is being incorporated into WebFOCUS, which is NCI's business intelligence software. In WebFOCUS, the application related to our model is called "Surgical Analytics." The retailer and developer of WebFOCUS (Information Builders) intends to make Surgical Analytics commercially available to all of its WebFOCUS users.

Following the model described in this paper, the primary role of Surgical Analytics is to provide ward occupancy projections for given MSS proposals. All data needed for the model is automatically extracted from the hospital information management systems. In addition to ward occupancy projections, Surgical Analytics provides additional information to hospital managers. This additional information and additional functionality of Surgical Analytics is discussed in the remainder of this section.

The Surgical Analytics package provides additional ways to evaluate MSS proposals, including:

- *The required number of nurses:* Hospitals typically staff wards according to a nurse-to-patient ratio. For example, on a medium care

ward a single nurse can care for three patients, thus if there are 12 patients on the ward then four nurses are needed. Using the ward occupancy projected by our model, the required number of nurses for a MSS proposal can be approximated using nurse-to-patient ratios.

- *Resource conflicts:* Some surgical procedures require specialized and expensive equipment of which there is a limited supply. This means that problems can occur when two specialties that typically use the same equipment are scheduled at the same time. As a result, certain MSS proposals are infeasible. The Surgical Analytics package allows resource conflicts to be indicated such that infeasible MSS proposals can be easily identified.

- *Scheduling conflicts:* Surgeons typically have commitments in areas of the hospital other than the OR. As a result, scheduling surgeons at certain times is not possible and thus certain MSS proposals are infeasible. The Surgical Analytics package allows physician schedules to be indicated such that infeasible MSS proposals can be easily identified.

Finding the best MSS (in terms of, for example, ward occupancy) within these scheduling and resource constraints is a prime area for further research. The advantages and disadvantages of using a search heuristic is discussed briefly in the Discussion section.

The Surgical Analytics package reports several performance metrics derived from the hospital's historical records. These metrics are not intended to help evaluate a MSS proposal but rather to provide some direction as to how the current MSS should be changed. Specifically, the metrics help to determine if the number of OR blocks assigned to each specialty is appropriate given, for example, each specialty's utilization, production and patient waiting times. Some of these metrics include:

- *OR hours used versus OR hours allocated:* This metric allows management to evaluate whether or not specialties are using less (or

more) OR hours than allocated to them. There can be a variety of reasons explaining why a specialty used less OR hours than they were allocated, including a shortage of patients, shortage of staff, surgeon vacations, etc. Depending on the reason(s), managers may allocate less OR time to such specialties when redesigning the MSS.

- *Completed cases versus production targets:* This metric allows management to evaluate whether or not specialties are on track to reach their production targets. Specialties not on track may be allocated additional OR time when redesigning the MSS.

- *Patient waiting time:* This metric allows management to see how long patients are waiting for surgeries. Surgical specialties with long waiting times may be allocated more OR time when redesigning the MSS.

6.5 Discussion

In many hospitals the surgical department has the foremost influence on the hospital's workload. As such, its activities (or lack thereof) cause a ripple effect elsewhere in the hospital. Upstream processes are less sensitive to changes, as there is often a waiting list for surgical operations which acts as a buffer, dampening the effect. For downstream process this is different, as a buffer of post-surgery patients waiting to be admitted to a ward cannot exist. Since post-surgical activities are sensitive to the activities in the OR, it is important to derive one in terms of the other. As described in this chapter, the workload for downstream departments can be described as a function of the MSS.

With the approach discussed in this chapter, a new MSS was developed for NCI which reduced the fluctuations in the daily ward census, creating a more balanced workload on the wards. The roll out of the new MSS corresponded with the opening of an additional OR which was expected to overwhelm the wards. By using the process described in this chapter to develop a MSS that accounted for the inpatient

wards, peaks in ward occupancy were reduced. As such capacity will be used more efficiently and the hospital has the means to support the additional OR without a major expansion in the wards.

The main benefit of the model was the ability to quantify the concerns of ward staff, thereby providing a platform from which they could begin to negotiate a solution. Staff was quick to embrace the model output, particularly after seeing several modifications to the original MSS, at which point they were able to roughly predict the model output intuitively. For example, on Thursdays and Fridays the wards tended to be crowded with patients. To remedy this, specialties that completed many cases per OR block were removed from Thursday and Friday OR blocks and assigned to OR blocks earlier in the week. To accommodate these changes, specialties which complete a relatively small number of cases per OR block were moved to Thursday and Friday. Once staff could foresee the impact of swapping one surgical OR block assignment with another, the implemented MSS came quickly.

In the model described in this chapter, we treated the equipment and physician schedule restrictions as unchangeable. It is possible that further improvements in the ward occupancy could have been achieved if these restrictions were relaxed. In this way the model can be used to illustrate the benefits of buying an extra piece of equipment or of changing physicians' schedules. An additional restriction, which if relaxed may have allowed further improvements, is the assignment of wards to surgical specialties. In other words, in addition to changing when a specialty operates, it may prove advantageous to change which ward the patients are admitted to after surgery. Furthermore, we chose the best MSS from those created through swapping OR block and surgical specialty assignments. It is possible that a search heuristic may have found a better MSS, although it would have required the many surgical department restrictions to be modelled and the more complex model may not have garnered the same level of staff understanding and support.

Finally, variants of this model have been applied in three other Dutch hospitals. At two of the hospitals the model was used to improve the

workload balance on the inpatient wards. In the third hospital it was used to evaluate the allocation of beds to wards.

Chapter 7

Pharmacy policies to reduce waiting times

Contents

7.1 Introduction

The chemotherapy day unit (CDU) at NCI is concerned about long waiting time for patients after they arrive at the hospital. Initial analysis indicated that a large percentage of this wait is caused by medication preparation process in the pharmacy. In this chapter we examine the process by which medication orders are placed by the CDU, and

filled by the pharmacy. More specifically, we evaluate whether a reduction in waiting time resulting from medication orders being prepared in advance of patient appointments is justified, given that medications prepared in advance are wasted when patients arrive too sick for treatment. We use operational research models to evaluate how changes in this process affect both pharmacy costs and patient waiting times. The relationship between the pharmacy and the CDU is described below.

Patients receiving chemotherapy do so over a number of months, with weekly or biweekly appointments. Each appointment is scheduled at least one week in advance. On the day of their appointment a patient either reports directly to the CDU or to the laboratory. Patients reporting to the laboratory require a blood test, which is used to assess if they are healthy enough (i.e. fit) to receive their scheduled course of chemotherapy. Patients who are fit receive the chemotherapy, those that are not are rescheduled for a later time. In this case study we found that approximately 80% of patients require a blood test and approximately 10% of them are found to be unfit to receive chemotherapy. Patients reporting directly to the CDU (i.e. patients not requiring a blood test) receive a quick health check before receiving chemotherapy treatment. Approximately 5% of these patients are found to be unfit and do not receive treatment.

Current practice states that chemotherapy medications are only to be prepared after the patient is deemed "fit" to receive treatment, and, is present in the CDU. This practice of preparing the medication "on demand" is motivated by the high cost of many of the medications. The pharmacy does not wish to prepare a medicine before they are sure the patient can receive it, as this ensures no medicines are wasted. They argue that since chemotherapy medications can cost up to €1800 per treatment, it is prudent to be sure they will be used before preparing them. Unused medicine may contribute considerably into the operational waste of a hospital [25].

Management of the CDU, on the other hand, argues that preparing medicines on demand adds an extra process step leading to unnecessary waiting for patients. CDU management would prefer if the

medications were prepared "in advance" of the appointment so that patients could receive their chemotherapy immediately after they have been found fit to receive it. They argue that the percentage of patients found to be unfit is sufficiently low to justify preparing medications in advance.

To determine which (if any) medication should be prepared in advance, a number of additional factors should be considered. Different medications have different shelf lives which can dictate how soon in advance of an appointment a medicine can be prepared. Different medications also have different costs; generic drugs are significantly cheaper than brand name drugs and those which are new. Some medications are also more toxic, which results in a higher percentage of unfit patients. Finally, some medications are used more frequently than others which allows them to be given to a different patient, should the original patient be found unfit. Because of these many factors and the uncertainty involved in this process, it is unclear to management which medicines should be prepared in advance and which medicines should be prepared on demand.

The purpose of this chapter is twofold. First, as a case study at NCI, we define a policy stating which medicines should be prepared in advance. This policy strikes a balance between the cost of wasted medicines and the 'cost' of waiting patients. The second purpose is to describe and evaluate an analytical model with explicit expressions that allows this analysis to be easily repeated at other hospitals.

Using operational research models in health care settings is not new (e.g. see [91, 165, 179]), however the majority of studies of outpatient clinics focus primarily on scheduling [39]. Improvements related to operational processes have seen less attention, and in particular, models considering the processes of two interacting departments are uncommon [75, 193]. Most closely related to the work presented in this chapter, is by van Merode et al. [144] where a system dynamics simulation is used to study cost and waiting times for a wide range of chemotherapy patient and drug types. Whereas our model is used to define a policy to guide daily decision making by pharmacists, their

model considers how different chemotherapy delivery protocols affect patient satisfaction and costs.

This chapter contributes to the literature of health care management science through the derivation of analytic expressions for patient waiting times and medication costs, within the described context. Although this context is a specific operational area, our approach provides an effective and readily implementable methodology for addressing the problem. The analytic expressions have the distinct advantage over simulation techniques in that changes in model parameters can easily be accounted for without needing to repeat model runs. Furthermore, specialized software and trained modellers are not needed to repeat the analysis when medication costs change or when new medications/protocols are introduced. In the same vein, other hospitals with a different patient case mix can easily complete the analysis in their setting without specific software or simulation know-how.

The chapter is organized as follows. The queueing system and model are introduced in Section 7.2, the waiting time analysis in Section 7.3, and the analysis of the cost of wasted medication orders in Section 7.4. The use of the model for policy decisions at NCI is discussed in Section 7.5, and a general discussion on the model's applicability to other other hospitals is discussed in Section 7.6.

7.2 Model description

The process introduced in the previous section is examined analytically and with simulation. In this section we describe the analytical model (which is an approximation) and the simulation model.

7.2.1 Model flow

The queueing system for medication orders in the pharmacy is described as follows: The system consists of two queues leading to a

server with c pharmacists. Orders O_t^1 (where t indexes time measured in days) go to the first queue and wait there until being prepared. These orders are started only after their corresponding patient is deemed fit and must be finished by the end of the day. Orders O_t^2 go to the second queue and are not required to be completed the same day. These are the orders for patients who will arrive at the hospital on the next day. At the end of each day orders that are still present in the second queue, join the first queue the next day. These orders are called the 'backlog' and are denoted by L_t. Orders in the first queue have (non-preemptive) priority over orders in the second queue, since their corresponding patients are waiting. O_t^1 orders are those that are completed when the patient arrives at the hospital (and is deemed fit for treatment) and O_t^2 orders are prepared in advance of the patient's arrival.

The queueing system for patients in the CDU is described as follows: Upon arrival it is determined if patients are fit to receive their scheduled treatment of chemotherapy. Unfit patients are sent home without receiving treatment. Fit patients whose medicine orders are part of the completed orders immediately receive treatment. Fit patients whose medicine orders are *not* part of the completed orders must wait until their medication order is complete before receiving treatment. Figure 7.1 depicts the flows of orders and patients.

The decision required in this problem is to determine which medicines should be denoted as O_t^1 orders, and conversely, which should be denoted as O_t^2 orders. This decision is evaluated on the resulting waiting time for patients (Queue 3 from Figure 7.1) and the expected cost of wasted medicines (medicines prepared in advance for patients who are later found to be unfit to receive their treatment).

Analytically the system cannot be described in a straightforward manner since it contains two time scales. Orders arrive during the entire day and their waiting times are measured in minutes. The backlog, on the other hand, occurs only at the end of the day, creating arrivals to Queue 1 on the next day.

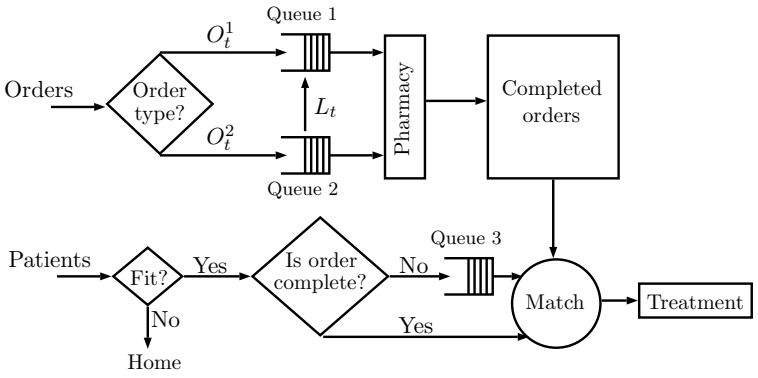


Figure 7.1 – The process model

In order to analyze the waiting time in this model analytically we split it into two submodels. The first submodel observes the process on a day-to-day level and allows us to determine the expected amount of backlog on a day. This submodel is described in Subsection 7.3.1. Using the expected amount of backlog, and given that the arrival rate to Queue 1 is known, the waiting times of the patients can be determined, as described in Subsection 7.3.2. An analytic expression to compute the expected cost of wasted medicines is described in Section 7.4.

The system is also modelled with a discrete event simulation programmed in MatLab. Each decision denoting which medicines are O_t^1 orders (and conversely, which are O_t^2 orders) represents a "what if" scenario in the simulation. For each scenario, we simulate 50 repetitions of 1000 days (we use 20 days as a warm up period) to determine the corresponding waiting time and wastage cost. This led to very tight confidence intervals for the model output (e.g. for average waiting times, the 95% confidence intervals had a width of less than 0.1 minutes) hence the confidence intervals are not reported in the text. The results from the simulation are used principly to evaluate the validity of the analytic expressions.

Both Sections 7.3 and 7.4 conclude with numerical results related to the NCI case study. In these subsections the simulation results are

compared to the results from the analytic model. The NCI case study data used for this purpose is introduced in the following subsection.

7.2.2 Model data

The total number of patients that arrive on a day t is denoted by N_t, $N_t \sim \text{Poisson}(\lambda)$ following from the data. In this case study $\lambda = 69.9$. A workday of the pharmacy consists of 555 minutes, starting at 08:15h and ending at 17:30h. The chemotherapy appointments begin in batches every fifteen minutes from 08:15h until 15:30h. Thus, the arrival of chemotherapy drug orders O_t^1 are spread over a period of $T = 450$ minutes. From the data, we found that the number of arrivals at a fifteen minute time slot τ has a $\text{Poisson}(q_\tau \cdot \lambda)$ distribution, where q_τ is an estimated fraction of patients that arrive at this time slot. Patients require a blood test with probability 0.8. Patients are found to be unfit for treatment with probability 0.1 in the case they had a blood test, and with probability 0.05, otherwise. This results in the fraction $r = 0.8 \cdot 0.1 + 0.2 \cdot 0.05 = 0.09$ of unfit patients. In our case study we found this fraction to be the same for each type of medicine, although in general it does not need to be.

Let the set of all 52 medicines be denoted by S. The set of medicines that are not allowed to be prepared in advance (i.e. are O_t^1 orders) is denoted by S_1. The set of medicines that are allowed to be prepared in advance (i.e. are O_t^2 orders) is denoted by S_2. We assume that each patient requires medicine $i \in S$ with probability p_i independently of other patients. It follows that the number of patients that need medicine i is $\text{Poisson}(\lambda \cdot p_i)$ distributed. The probability p_i is determined as follows,

$$p_i = \frac{f_i}{\sum_{j \in S} f_j},$$

where f_i is the number of times medicine i was used in a year and varies between 1 and 2941. These numbers follow from the data of the hospital. The price of a medicine i is denoted by c_i and ranges from approximately €1 to €1800 per order.

The preparation times of medicines are independent and identically distributed (IID) and are denoted by B, which is uniformly distributed between 5 and 20 minutes, as per the estimation of the pharmacists. The pharmacy is staffed by $c = 2$ pharmacists. Medicine orders arrive at the pharmacy 24 hours before the appointment time of that patient. The total number of orders that arrive on day t at the pharmacy is denoted by O_t.

7.3 Patient waiting times

7.3.1 Backlog

In this subsection we describe a slotted queueing model to calculate the expected amount of backlog. Let L_t be the amount of backlog on day t, this L_t depends on the number of orders O_t^1 and O_t^2, the amount of backlog on the previous day L_{t-1} and the capacity (in terms of orders) of the pharmacy K_t. The capacity of the pharmacy is the maximum number of orders that can be prepared in a single day. This slotted model, where one slot equals one order, does not take the arrival time of the patients into account, only the orders, which arrive in a batch the day before the patient arrives. It is clear that the sum of L_{t-1}, O_t^1 and O_t^2 equals the total number of orders that need to be handled on day t, so the following equation arises,

$$L_t = (L_{t-1} + O_t^1 + O_t^2 - K_t)^+, \tag{7.1}$$

where $x^+ = x$ if $x \geq 0$ and $x^+ = 0$ otherwise. Note that since each order corresponds to one patient then,

$$O_t^1 + O_{t-1}^2 = N_t.$$

Assume that on each day $\sum_{i \in S_2} p_i$ orders may be prepared in advance. Then O_{t-1}^2 and O_t^2 are identically Poisson distributed with parameter $\lambda \cdot \sum_{i \in S_2} p_i$. Following the model assumptions O_t^2 is independent of

O_t^1. Since the expected number of patients is the same each day, the following holds,

$$O_t^1 + O_t^2 = O_t \stackrel{d}{=} N_t.$$

Furthermore we make the assumption that the O_t is independent of K_t, then equation (7.1) can be written as,

$$L_t = (L_{t-1} + O_t - K_t)^+. \tag{7.2}$$

Equation (7.2) is known as Lindley's recursion (see e.g. [46]). A way to approach a solution of Lindley's recursion is by directly solving the corresponding Markov chain. The states of this Markov chain are defined by the number of backlog orders. The transition probabilities of this Markov chain are given by,

$$P_{ij} = \begin{cases} \sum_{k=i}^{\infty} \mathbb{P}(O_t \leq k-i)\mathbb{P}(K_t = k) & \text{if } j = 0, \\ \sum_{k=i-j}^{\infty} \mathbb{P}(O_t = j-i+k)\mathbb{P}(K_t = k) & \text{if } 0 < j \leq i, \\ \sum_{k=0}^{\infty} \mathbb{P}(O_t = j-i+k)\mathbb{P}(K_t = k) & \text{if } j > i. \end{cases}$$

Since O_t is Poisson distributed, we have,

$$P_{ij} = \begin{cases} \sum_{k=i}^{\infty} \sum_{j=0}^{k-i} \frac{e^{-\lambda}\lambda^{j-i+k}}{(j-i+k)!} \mathbb{P}(K_t = k) & \text{if } j = 0, \\ \sum_{k=i-j}^{\infty} \frac{e^{-\lambda}\lambda^{j-i+k}}{(j-i+k)!} \mathbb{P}(K_t = k) & \text{if } 0 < j \leq i, \\ \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^{j-i+k}}{(j-i+k)!} \mathbb{P}(K_t = k) & \text{if } j > i. \end{cases}$$

Following from renewal theory, we model K_t as a normally distributed random variable. Following from the data, K_t has mean $\mu = 87.92$ and variance $\sigma^2 = 10.66$. Renewal theory results imply that the number of IID random variables (preparation times) that can be fitted in a large time interval (working day of the pharmacy) is approximately normally distributed with mean and variance defined by the first three moments of the preparation times, see e.g. Ross [166, Chapter 3]. In this case

$\mathbb{P}(K_t = k)$ is approximated by $\mathbb{P}(k - 0.5 < K_t \leq k + 0.5)$. To solve the Markov chain, the chain is considered to be finite with N states, where N is sufficiently large such that the probability to transit to states larger than N is negligible. The expected number of backlog orders ($\mathbb{E}[L]$) follows from the steady state distributions.

7.3.2 Waiting times

Knowing the amount of backlog, the total number of orders in the priority queue (Queue 1) can be determined and the waiting times can then be approximated. We explain the approximation in the following three steps. In the first step, the load on the system resulting from the different order types is determined. In the second step, we explain how the batch arrival process can be approximately modelled by IID random variables. This adaptation of the arrival process makes it possible to use existing waiting time approximations to solve our problem. The approximation chosen for our purpose is explained and formulated in the third step.

Let the expected waiting time of patients that receive a medicine from set S_1 be denoted by $\mathbb{E}[W_{S_1}]$. This includes the time waiting for the pharmacist to begin preparing the order and also the preparation time. This is the only waiting time of concern in our problem as the patients corresponding to these medicine orders are already present in the hospital.

System load: The pharmacy workload results from the two order types, O_t^i, $i = 1, 2$ (see Figure 7.1). Let λ_i be the average number of new orders for each i arriving in one day (this excludes possible backlog from the day before). The orders of type O_t^1 are prepared by the pharmacy on day t, only if the patient is found fit for treatment which happens with probability $1 - r$. As many O_t^2 orders as possible, within the opening hours of pharmacy, are completed on day t. Those orders which are not completed due to a shortage of time form the backlog L and are added to λ_1 on day $t + 1$. Furthermore, all orders arrive during the time period when appointments take place, i.e. time period

T . Thus, let ρ denote the amount of work offered to the system per minute, then we can write,

$$\rho = \rho_1 + \rho_2,$$

where,

$$\rho_1 = \mathbb{E}[B](\lambda_1 + \mathbb{E}[L])(1-r)/T, \quad \rho_2 = \mathbb{E}[B]\lambda_2/T$$

and B is the preparation times of the medicines as introduced in Subsection 7.2.2.

Arrival process: We evaluate $\mathbb{E}[W_{S_1}]$ using the approximating formulas for the $GI/G/c$ priority queue. Note however, that the inter-arrival times in our model are not IID because patients arrive in batches. To remedy this, we use two approximation steps. First, we assume that the fraction q_τ of patients arriving at slot τ is the same for each $\tau = 1, \ldots, T/d$, where d is the time interval between two appointment slots. Then, the number of orders arriving at each slot becomes Poisson(λ_{slot}), with,

$$\lambda_{slot} = (\lambda_1 + \mathbb{E}[L])(1-r)d/T.$$

Next, we apply the so-called Equivalent Random Method. The idea behind this method is to replace a non-IID sequence of random variables by IID random variables with the same mean and variance, see [200] for classical references and [128] for recent applications in health care. Thus, we need to compute the mean and variance of inter-arrival times in our system and substitute these numbers into the approximation for the $GI/G/c$ queue.

Let $A_1, A_2, \ldots$ be inter-arrival times between two subsequent patients. Obviously, $\mathbb{E}[A] = d/\lambda_{slot}$. Next, observe that the inter-arrival time between two patients arriving at the same slot is zero. Hence, the renewal theory argument gives that,

$$\mathbb{P}(A > 0) = \frac{\mathbb{E}[\#\text{ non-empty batches in one slot}]}{\mathbb{E}[\#\text{ patients in one slot}]} = \frac{1 - e^{-\lambda_{slot}}}{\lambda_{slot}}.$$

Further, note that if $k-1$ slots are empty then the inter-arrival time between two non-empty slots becomes kd. Thus, we derive,

$$\mathbb{E}[A^2] = \frac{1-e^{-\lambda_{slot}}}{\lambda_{slot}} \sum_{k=1}^{\infty} (kd)^2 e^{-(k-1)\lambda_{slot}} (1-e^{-\lambda_{slot}}) = \frac{d^2}{\lambda_{slot}} \frac{1+e^{-\lambda_{slot}}}{1-e^{-\lambda_{slot}}}.$$

Define $c_X^2 = \mathrm{VAR}(X)/(\mathbb{E}[X])^2$ as the squared coefficient of variance of random variable X. Then the calculations above give,

$$c_A^2 = \frac{\mathbb{E}[A^2]}{(\mathbb{E}[A])^2} - 1 = \frac{\lambda_{slot}(1+e^{-\lambda_{slot}})}{1-e^{-\lambda_{slot}}} - 1.$$

Waiting time: We now provide the approximation for the waiting time in different scenarios. First we consider $S_1 = S$ which means that no medicines are prepared in advance. Thereafter S_1 is considered to be a subset of S such that $S_2 = S \backslash S_1$ is non-empty, i.e. some medicines are prepared in advance.

In case $S_1 = S$ we have a queueing system with c servers (pharmacists) and IID service times. The number of arrivals per day is $\lambda = \lambda_1$, and by definition in this case $\mathbb{E}[L] = 0$. We evaluate the waiting time of the patients in this queueing system using the approximation from [201] for the average waiting time $\mathbb{E}[W(GI/G/c)]$ in the $GI/G/c$ queue,

$$\begin{aligned} \mathbb{E}[W_{S_1}] &= \mathbb{E}[W(GI/G/c)] + \mathbb{E}[B] \\ &\approx \left(\frac{c_A^2 + c_B^2}{2}\right) \mathbb{E}[W(M/M/c)] + \mathbb{E}[B] \\ &= \left(\frac{c_A^2 + c_B^2}{2}\right) \mathbb{E}[B] c^{-1} \cdot D/[1 - c^{-1} \cdot \rho] + \mathbb{E}[B], \quad \text{if } S_1 = S \end{aligned} \tag{7.3}$$

where c_B^2 is the squared coefficient of variance of the service time and where D denotes the probability of delay in a $M/M/c$ queue. The expressions for $\mathbb{E}[W(M/M/c)]$ and D can be found e.g. in Tijms [185], the latter formula being given by,

$$D = \frac{\rho^c}{c!} \left(\left(1 - \frac{\rho}{c}\right) \sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!} \right)^{-1}.$$

In the case where $S_1 \subset S$ such that $S_2 = S \backslash S_1 \neq \emptyset$, a priority rule is used. The medicines that are prepared for patients on that same day have priority over the medicines that are prepared for the next day. Because pharmacists cannot shelve unfinished medicine and resume preparation later, a non-preemptive priority system is used. For this priority system we use the approach of Federgruen and Groenevelt [73], which gives the next approximation formula for the waiting time in the priority queue,

$$\mathbb{E}[W_{S_1}] \approx \left(\frac{c_A^2 + c_B^2}{2} \right) \mathbb{E}[B] c^{-1} \cdot D / [1 - c^{-1} \cdot \rho_1] + \mathbb{E}[B], \quad \text{if } S \backslash S_1 \neq \emptyset. \tag{7.4}$$

The numerical results in the next section prove that our analytical model gives a good approximation of the expected waiting times.

7.3.3 Numeric results

Given that there are S medications, and that each could be prepared in advance, then there are $S!$ different policies to evaluate. However, in the interest of having an easily implementable policy, NCI decided that a simple criterion to identify "prepare in advance" medicines (i.e. O_t^2 orders) should be used. To this end, the price of the medicine is used as the criterion to indicate which medicines should be prepared in advance. Specifically, if the price of a medicine is less than M euros, then the medicine is to be prepared in advance. This limits the number of policies to evaluate to at most, S. When $M = 0$ no medicine orders are prepared in advance, and when $M = 1000$ all medicine orders (which have shelf life greater than 24 hours and are not extremely expensive) are prepared in advance. The most expensive medicines (i.e. medicines which cost more than €1000 per dose) are not considered candidates to be prepared in advance. Should a hospital choose a different criterion for defining their policy, the analysis of Subsection 7.3.2 remains valid. The numeric results presented in this chapter are limited to only 9 policies, but include both border policies, i.e. $M = 0$ and $M = 1000$.

Amount prepared in advance:	M	λ_1	$\mathbb{E}[L]$ (simulation)	$\mathbb{E}[W_{S_1}]$ (simulation)	$\mathbb{E}[W_{S_1}]$ (analytical)
None	0	69.9	0.00	46.1	55.9
	10	66.1	0.01	42.3	42.2
	20	57.0	0.03	33.7	28.6
	30	46.4	0.05	26.3	22.4
	40	34.7	0.07	21.1	19.2
	100	22.2	0.08	18.0	17.3
	200	18.7	0.09	17.3	17.0
	500	18.2	0.09	17.2	16.9
All*	1000	6.3	0.09	15.6	16.1

*All medicines which have shelf life greater than 24 hours and are not extremely expensive

Table 7.1 – Waiting times for different policies M

Table 7.1 shows for a chosen M the resulting λ_1, the amount of backlog and the expected waiting times of patients. The waiting times computed analytically and with simulation are displayed.

The expected volume of backlog increases as more medicines are prepared in advance. This is in line with the observation that if none of the medicines are prepared in advance then there are no medicines in the second queue and there will be no backlog. In all cases, patients only have to wait if their medicine was not allowed to be prepared in advance, or if their order was backlogged at the end of the previous day.

It is also not surprising that waiting times are smallest in the cases where most medicines are prepared in advance. In this case only those people requiring a medication with a shelf life of less than 24 hours (or an extremely expensive medication) wait for their medicine to be prepared. The results show clearly that preparing more medicines in advance results in lower waiting times. The difference between the numerical and analytical results is due to the two approximations in (7.3). The first approximation replaces the original batch arrival

process with IID arrivals, which may result in an error of a couple of minutes. We refer to the comments on formula (70) in [201] for the conditions under which the approximation for $\mathbb{E}[W(GI/G/c)]$ is sufficiently precise.

7.4 Cost of wasted medicine

The downside of preparing medicines in advance is that these orders risk being wasted, resulting in additional costs for the pharmacy. In this section we formulate an expression to compute the expected cost per day $\mathbb{E}[C]$ of wasted medicines for a given Policy M. First, we assume that all medicines prepared in advance are wasted when the corresponding patient is found to be unfit for treatment. In Subsection 7.4.2 we assume that these orders can be stored and later given to a different patient.

The expected cost per day $\mathbb{E}[C]$ can be calculated in straightforward manner. Recall that S_2 denotes the set of medicines that are allowed to be prepared in advance and let c_i be the price of medicine i and r_i the probability that a treatment with medicine i is wasted due to a patient being found unfit for treatment. Then the following holds,

$$\mathbb{E}[C_{order}] = \sum_{i \in S_2} c_i \cdot r_i \cdot p_i + g \cdot r_i \tag{7.5}$$

where g is the non-material cost to pharmacy to prepare an order (i.e. overhead costs and staff wages). In the NCI case study $g = 0$ (without loss of generality) since there was no intention to reduce the number of pharmacy staff and therefore g was considered an irrelevant sunk cost.

Knowing the cost per order and the frequency of orders, the expected daily cost can be computed as follows,

$$\mathbb{E}[C] \quad = \quad \mathbb{E}[C_{order}] \cdot \mathbb{E}[N_t] = \sum_{i \in S_2} c_i \cdot r_i \cdot p_i \cdot \lambda. \tag{7.6}$$

Amount prepared in advance:	M	$\mathbb{E}[C]$ (simulation)	$\mathbb{E}[C]$ (analytical)
None	0	0.0	0.0
	10	1.9	1.9
	20	11.3	11.3
	30	35.4	35.4
	40	71.2	71.1
	100	124.9	124.9
	200	167.6	167.5
	500	185.4	185.7
All*	1000	897.1	900.0

*All medicines which have shelf life greater than 24 hours and are not extremely expensive

Table 7.2 – Expected daily cost of wasted medicines for different policies M

7.4.1 Numeric results

In Table 7.2 the expected costs computed by (7.6) and by simulation are shown for various policies M. From the table, it is clear that when all medicines are prepared in advance, the costs are very high. These high costs motivate us to analyze storing the medicines from cancelled treatment to "reused" by other patients.

7.4.2 Reuse of medicines

Reusing medicines causes fewer to be wasted and therefore reduces the cost for pharmacy. However, given the complications (and possible risks) associated with managing an inventory of "to be reused" medicines, it is likely that only expensive and frequently used medicines will be stored for later use and the others will be discarded.

Let $\mathbb{E}[C_i]$ be the expected daily wastage cost of the medicine i where,

$$\mathbb{E}[C_i] = c_i \cdot r_i \cdot p_i \cdot \lambda. \tag{7.7}$$

To investigate the effect of reusing medicines, we introduce policy F. Let policy F indicate which medicines can be stored for reuse (and conversely, which cannot be stored for reuse) after a treatment is cancelled. We define F such that if the $\mathbb{E}[C_i]$ of medicine i is higher than F, then that medicine can be stored for reuse (should a hospital choose a different criterion for indicating which medicines can be stored for reuse, the following analysis remains valid). Let S_F denote the set of medicines that are to be stored for reuse. In calculating the expected cost resulting from policy F we make the assumption that S_F medicines are *never* wasted, and compute the cost as follows,

$$\mathbb{E}[C_{order,F}] = \sum_{i \in S_2 \setminus S_F} c_i \cdot r_i \cdot p_i + g \cdot r_i \tag{7.8}$$

Note that the expected daily cost associated with policy F is computed with (7.6) by substituting $\mathbb{E}[C_{order}]$ with $\mathbb{E}[C_{order,F}]$.

To test the validity of our assumption that S_F medicines are *never* wasted, we use a Markov model. Consider medicine i in S_F which has a shelf life of two days, that is, it is prepared one day before a planned treatment and can be used either on the day of the treatment (t) or the day after treatment. For this medicine we describe the state of the Markov chain with vector (x, y) where $x + y$ is the number of unused orders of the medicine in storage at the end of day t (x orders can be reused the day after, and y orders cannot be reused and thus are wasted). The transition probabilities from state (a, b) to state (c, d) depend only on the number of orders for day $t + 1$ and can be found explicitly. Using this model, the probability of wasting a stored medicine was found to be negligible. The same observation followed from the simulation runs for the range of policies considered in the NCI case study.

Alternatively, in other settings, the cost of wasted medicines computed by (7.8) should be considered the best case (i.e. a lower bound of the cost). Intuitively, note that a medicine with a two days shelf life prepared for day t is wasted only if the number of fit patients requiring the medicine on day $t+1$ is smaller than the number of unfit

Amount prepared in advance:	M	F 300	100	50	20	10	5	1
None	0	0	0	0	0	0	0	0
	10	1.9	1.9	1.9	1.9	1.9	1.9	0.6
	20	11.3	11.3	11.3	11.2	11.3	11.3	1.5
	30	35.4	35.4	35.4	35.4	35.4	20.8	2.3
	40	71.1	71.1	71.1	71.1	44.4	22.8	2.3
	100	124.9	124.9	124.9	79.2	52.5	30.9	3.6
	200	167.5	167.5	167.5	94.4	67.6	36.2	3.6
	500	185.7	185.7	185.7	112.5	73.7	42.3	4.2
	1000	371.1	252.4	188.0	114.9	76.1	44.6	4.2
All*	2000	705.2	290.2	225.9	127.8	76.1	44.6	4.2

*All medicines which have shelf life greater than 24 hours

Table 7.3 – Expected daily cost of wasted medicines for different policies M and F

patients requiring the medicine on day t. Furthermore, the probability of wasting a medicine is even smaller for medicines with a shelf life of more than two days. Hence, if r_i is small relative to the number of new orders for medicine i (as in our case study) then the event of wasting a medicine is highly unlikely. Therefore, the lower bound (7.8) gives a good indication of wastage costs for a wide range of scenarios.

Finally reused medicines require approximately the same amount of preparation time by the pharmacists (e.g. to check or adjust the dosage) and therefore the waiting time for patients is independent of F. Table 7.3 shows the expected cost for the pharmacy for a variety of polices M and F.

7.5 Application

The numeric results in Tables 7.1 and 7.2 illustrate how the decision to prepare medicines in advance influences both the waiting time for

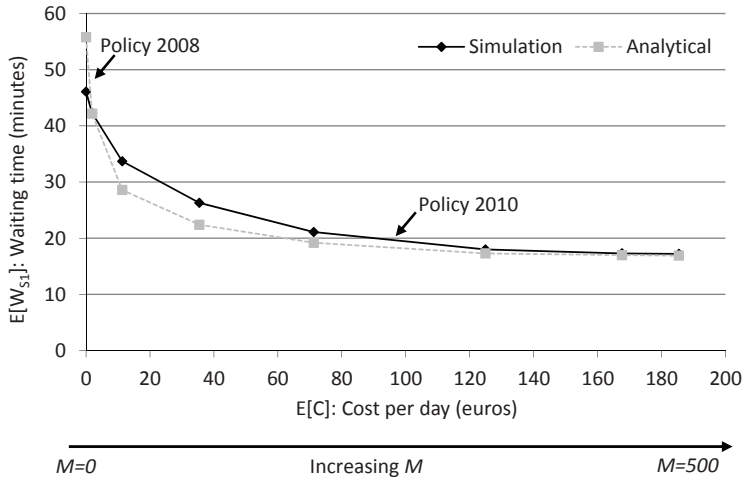


Figure 7.2 – Costs and waiting times for different policies M

patients and the costs for pharmacy. We represented this multiple criteria decision [204] graphically (see Figures 7.2 and 7.3) to allow management to see the relative advantages and disadvantages of each policy decision. By showing this tradeoff between the waiting time for patients and the cost for pharmacy, the hospital was able to make an informed decision.

In September of 2009, after considering the results presented in this chapter, management from the pharmacy and the CDU agreed that the shorter waiting times justified preparing certain medicines in advance. Furthermore, they chose to reuse their most expensive medicines. Based on this research the hospital is currently preparing approximately 80% of all medicines in advance. In this section we discuss and highlight the improvements resulting from this policy change.

To compare different values of M independent of F, Figure 7.2 plots M as a function of both waiting times and costs. The results from both the simulation and analytic approach are shown. The policies of the hospital in 2008 and 2010 are also shown.

Each point on Figure 7.2 represents a different policy M. In general, results from the analytical model give a lower estimate of the waiting times. However the analytical results are close to those obtained by simulation when preparing multiple medicines in advance. The figure shows clearly that the new policy (Policy 2010) decreases the expected waiting time from 45 minutes to 23 minutes at a cost of €105 per day. It is important to realize that waiting patients occupy a place in the CDU, which is an indirect capacity loss. In our case study, a gain of 20 minutes per patient saves approximately 23 person-hours of waiting at the CDU each day. This capacity can be used for handling more patients, thus yielding obvious benefits for the hospital and/or ensuring a high service level such that the patients receive their appointments without delay. On the other hand, the cost borne by the pharmacy is reasonable, considering that the pharmacy prepares approximately €8,000 worth of medicines each day, meaning the new policy accounts for only a 1-2% percent increase in its costs.

Figure 7.3 also plots M as a function of both waiting times and costs. The difference, however, is that in Figure 7.3, we consider a policy $F \approx 20$ such that the most medicines are reused.

Figure 7.3 shows clearly that lower costs can be achieved by reusing some medicines. Specifically in this case, a waiting time of 23 minutes can be achieved at a cost of €70 per day. This is €35 cheaper than in the case where no medicines are reused.

7.6 Discussion

Initially this system at NCI was analyzed with a simulation model. This approach was chosen due to the multiple time scales and because hospital staff were more familiar with this approach. However, when the case study component of this work was completed and it became apparent that improvements would be realized, we sought to formulate an analytical solution which would be easily reproduced in other settings or when prices of medicines change. As such, other hospitals

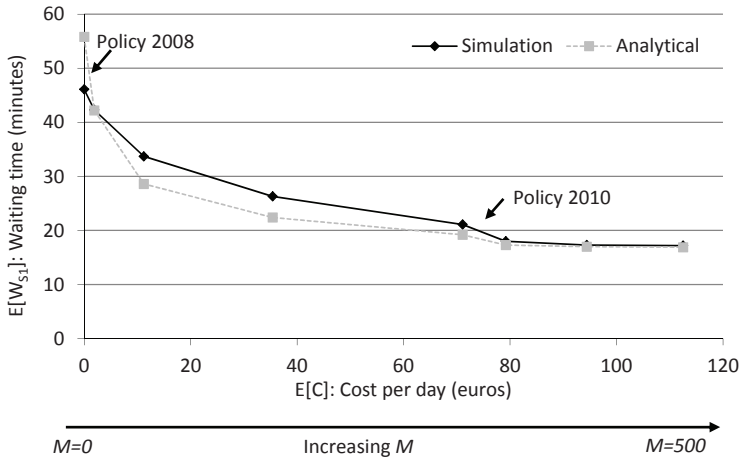


Figure 7.3 – Costs and waiting times for different policies M, with the reuse of expensive medicines ($F \approx 20$)

with similar practices as NCI can use the same methods to determine the waiting times of their patients and the cost to the pharmacy.

In case no medicines are prepared in advance, equation (7.3) gives an approximation of the waiting times. If a hospital has a policy of preparing multiple medicines in advance, equation (7.4) should be used. Equation (7.6) gives the cost to the pharmacy for each policy. The plots in Figures 7.2 and 7.3 can easily be reproduced in other hospitals to illustrate the costs and waiting times associated with various policies M and F (and likewise for the current policy). This approach leaves the decision autonomy with hospital managers, allowing them to decide how the waiting times of the patients and the cost of the pharmacy should be balanced (i.e. which metric should be given more weight).

The interactive nature of these two departments is a prime area for further study. The two departments have different and sometimes competing objectives and models, such as the one presented in this chapter, are needed to quantify the concerns of both. As shown, sig-

nificant improvements in patient waiting times can be gained at a modist cost for the hospital.

A topic for future research could be to examine possible implications of our research on appointment scheduling at the CDU. In this chapter we considered the appointment schedule as given. However, one can imagine that changes in the appointment schedule will affect the workload at the pharmacy. On the other hand, the pharmacy could develop a medicine preparation policy based on the appointment times of the patients at the CDU, for instance preparing medicines in advance only for patients that are scheduled in the morning.

In this analysis, M is defined such that once we indicate that a medicine is to be prepared in advance, all orders for this medicine are prepared in advance. It is likely that further improvements are possible if we relax this restriction and further specify this policy. For example, we could choose which orders to prepare in advance based on all outstanding orders by taking the length of the queue into account. This more complicated situation would better reflect the current state of the system, however it would likely require a model in real-time. How to incorporate a real-time policy into current practice, and examining if the gains justified the more complex policy, are additional areas for further research.

Chapter 8

Conclusion and outlook

In the introduction to this thesis, we argued that decisions made in health care are often made without explicit consideration for their impact on the whole system, i.e. the management style resembles a *reductionist approach* instead of a *systems approach*. We claim that this is in response to two adverse, but common, characteristics of health care. The first is complexity and the second is uncertainty. The research in this thesis illustrates example problems where these characteristics were present and provides solution approaches and general results to support decision makers coping with complexity and uncertainty. In this way, the research supports hospitals transitioning from a reductionist, to a systems approach to management.

To conclude this thesis, we discuss how the models of Chapters 3 and 4 aid in *reducing* complexity and uncertainty and how the models of Chapters 5, 6, and 7 aid in *coping* with complexity and uncertainty. Furthermore, we make conjectures regarding promising future areas of research related to these topics.

Reducing complexity and uncertainty: Chapters 3 and 4 analyze strategic decisions that can lead to less complex and more predictable hospital environments. Both chapters relate to limiting the range of treatments and services (i.e. specialization) offered by the hospital. By

supporting decisions to achieve a specific (or specialized) case mix, the model of Chapter 3 can help hospitals specialize. By analyzing the effects of pooling resources by patient diagnosis (i.e. creating focused factories to provide specific treatments to a homogeneous patient population), the model of Chapter 4 can help a hospital become more specialized. Through limiting the range of treatments, complexity is decreased and the patient population becomes more homogeneous (i.e. less uncertain). Given that hospitals are notoriously complex and unpredictable, and since specialization can remedy this (to some extent), hospitals are well motivated to specialize.

Hospitals have other incentives to specialize as well, in particular fee-for-service financing. With fee-for-service financing comes a motivation for hospitals to offer certain treatments through specialization. There are numerous examples. Hospitals that are for-profit will likely specialize in the most profitable services. Hospitals that are not-for-profit may specialize and use a profit in some services to fund other, less profitable, services. Hospitals may specialize in the most profitable services to fund innovation. Determining which services are "most profitable" for a given hospital depends on the expertise and infrastructure of the given hospital.

Looking to the future, hospital actions related to specialization offer potential opportunities for future research. From the hospital perspective, models such as that of Chapter 3 can support management in determining to what extent they should specialize. From a regional or national perspective (where population health is the objective), operations research models can be used to proactively identify service short comings and ensure that specialization of hospitals does not exclude important services. As such, exploring the benefits and disadvantages of specialization is a promising area for further research.

Coping with complexity and uncertainty: The problems addressed in Chapters 5, 6, and 7 are complex managerial problems that are plagued with uncertainty. Chapter 5 addresses a single departmental problem over a long time scale and Chapters 6 and 7 address multiple department problems over a shorter time scale. All three problems are

examples of using operations research models to solve complex and uncertainty problems.

Methodologically, the three models used in these chapters are different from each other. For example, in Chapter 5, queueing network models are used, in Chapter 6, the underlying model uses binomial distributions to model recovering patients, and in Chapter 7 simulation and approximations from queueing theory are used. Hence, using various techniques to cope with complexity and uncertainty is possible and often necessary.

Looking to the future, the thematic problem of Chapter 5 (namely that patients with chronic diseases require ongoing care for many years), is a promising area for further research. Since the prevalence of patients with chronic diseases increases as people live longer, more hospital resources are being used to care for people instead of curing people. If not appropriately controlled, this creates system congestion. As such, future health care logistics research should focus on system congestion and evaluate its impact on the overall performance of the hospital.

In addition to coping with complexity and uncertainty, the problems addressed in Chapters 6 and 7 are examples of the expanding scope taken by hospital decision makers. In both cases we see how a decision made by one department influences the operations of another. The models capture this interaction and quantify the effect of the decision on both departments, i.e. the models allow decision makers to consider the influence of their decisions on multiple components of the system. They still represent only a part of the entire system, however the scopes of the presented models are larger than that of many models presented previously in literature, and represent a step toward managing hospitals with a systems approach.

Looking to the future, the managerial style of hospitals will likely continue toward the systems approach. As a result, further modelling of the interactions of multiple departments will be a promising area for further research. Solving specific problems, as is done in Chapters 6 and 7, but also deriving more general results and frameworks,

is necessary.

Complexity and uncertainty in hospital processes is a key barrier inhibiting managers from using a systems approach. Operations research scientists can help eliminate this barrier by analyzing strategies to reduce complexity and uncertainty (e.g. by specialization) and by supporting managers in making sound decisions in the presence of complexity and uncertainty. The body of work in this thesis not only confirms that hospitals are very complex and that patient flows are uncertain, but also that operations research models are well equipped to deal with these characteristics and to evaluate initiatives to reduce them.

Bibliography

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

[2] Akkerman, R., Knip, M.: Reallocation of beds to reduce waiting time for cardiac surgery. Health Care Management Science **7**(2), 119–126 (2004)

[3] Allen, A.O.: Probability, Statistics and Queueing Theory. Academic Press, London (1990)

[4] Altinel, I.K., Ulas, E.: Simulation modeling for emergency bed requirement planning. Annals of Operations Research **67**(1), 183–210 (1996)

[5] Anthony, R.N.: Planning and control systems: a framework for analysis. Harvard Business School Division of Research, Boston (1965)

[6] Ashton, R., Hague, L., Brandreth, M., Worthington, D.J., Cropper, S.: A simulation-based study of a NHS walk-in centre. Journal of the Operational Research Society **56**(2), 153–161 (2005)

[7] Ata, B., van Mieghem, J.A.: The value of partial resource pooling: Should a service network be integratedor product-focused? Management Science **55**(1), 115–131 (2009)

[8] Bagust, A., Place, M., Posnett, J.W.: Dynamics of bed use in accommodating emergency admissions: stochastic simulation model. British Medical Journal **319**(7203), 155–158 (1999)

[9] Bailey, N.T.J.: A study of queues and appointment systems in hospital outpatient departments,with special reference to waiting times. Journal of the Royal Statistical Society **14**(2), 185–99 (1952)

[10] Balasubramanian, H., Banerjee, R., Denton, B., Naessens, J., Stahl, J.: Improving clinical access and continuity through physician panel redesign. Journal of General Internal Medicine **25**(10), 1109–1115 (2010)

[11] Balasubramanian, H., Denton, B., Lin, M.: Handbook of Healthcare Delivery Systems, chap. Managing physician panels in primary care, pp. 1–23. Taylor and Francis, New York (2010)

[12] Baskett, F., Chandy, K.M., Muntz, R.R., Palacios, F.G.: Open, closed, and mixed networks of queues with different classes of customers. Journal of the Association for Computing Machinery **22**(2), 248–260 (1975)

[13] Beland, F., Bergman, H., Lebel, P., Clarfield, A.M., Tousignant, P., Contandriopoulos, A.P., Dallaire, L.: A system of integrated care for older persons with disabilities in Canada: Results from a randomized controlled trial. Journals of Gerontology Series A: Biological and Medical Sciences **61**(4), 367–373 (2006)

[14] Beliën, J., Demeulemeester, E.: Building cyclic master surgery schedules with leveled resulting bed occupancy. European Journal of Operational Research **176**(2), 1185–1204 (2007)

[15] Beliën, J., Demeulemeester, E., Cardoen, B.: Visualizing the demand for various resources as a function of the master surgery schedule: A case study. Journal of Medical Systems **30**(5), 343–350 (2006)

[16] Beliën, J., Demeulemeester, E., Cardoen, B.: A decision support system for cyclic master surgery scheduling with multiple objectives. Journal of Scheduling **12**(2), 147–161 (2009)

[17] Bitran, G.R., Morabito, R.: Open queueing networks: Optimization and performance evaluation models for discrete manufacturing systems. Production and Operations Management **5**(2), 163–193 (1996)

[18] Black, D., Pearson, M.: Average length of stay, delayed discharge, and hospital congestion: A combination of medical and managerial skills is needed to solve the problem. British Medical Journal **325**(7365), 610–611 (2002)

[19] Blake, J.T.: Shooting arrows in the dark: the policies and practices of waitlist management in Canada. Clinical and investigative medicine **28**(6), 308–311 (2005)

[20] Blake, J.T., Carter, M.W.: Surgical process scheduling: a structured review. Journal of the Society for Health Systems **5**(3), 17–30 (1997)

[21] Blake, J.T., Carter, M.W., Richardson, S.: An analysis of emergency room wait time issues via computer simulation. INFOR **34**(4), 263–273 (1996)

[22] Blake, J.T., Donald, J.: Mount Sinai Hospital uses integer programming to allocate operating room time. Interfaces **32**(2), 63–73 (2002)

[23] Blasak, R.E., Armel, W.S., Starks, D.W., Hayduk, M.C.: The use of simulation to evaluate hospital operations between the emergency department and a medical telemetry unit. In: Proceedings of the 2003 Winter Simulation Conference, vol. 2, pp. 1887 – 1893 (2003)

[24] Blatchford, O., Capewell, S.: Emergency medical admissions: taking stock and planning for winter. British Medical Journal **315**(7119), 1322–1323 (1997)

[25] Boat, T.F., Chao, S.M., O'Neill, P.H.: From waste to value in health care. The Journal of the American Medical Association **299**(5), 568–571 (2008)

[26] Bonvissuto, C.A.: Avoiding unnecessary critical care costs. Healthcare Financial Management **48**(11), 47–52 (1994)

[27] Boucherie, R.J., Taylor, P.: Transient product from distributions in queueing networks. Discrete Event Dynamic Systems **3**(4), 375–396 (1993)

[28] Bowers, J., Mould, G.: Concentration and the variability of orthopaedic demand. Journal of the Operational Research Society **53**(2), 203–210 (2002)

[29] Bowers, J., Mould, G.: Ambulatory care and orthopaedic capacity planning. Health Care Management Science **8**(1), 41–47 (2005)

[30] Brailsford, S.C., Lattimer, V.A., Tarnaras, P., Turnbull, J.C.: Emergency and on-demand health care: modelling a large complex system. Journal of the Operational Research Society **55**(1), 34–42 (2004)

[31] Brasted, C.: Ultrasound waiting lists: rational queue or extended capacity? Health Care Management Science **11**(2), 196–207 (2008)

[32] Calichman, M.V.: Creating an optimal operating room schedule. Association of peri Operative Registered Nurses (AORN) Journal **81**(3), 580–588 (2005)

[33] Capewell, S.: The continuing rise in emergency admissions. British Medical Journal **312**(7037), 991–992 (1996)

[34] Cardoen, B., Demeulemeester, E., Beliën, J.: Operating room planning and scheduling: List of references (2008). Retrieved October 10, 2008, from http://www.econ.kuleuven.be/public/ndbaa92/bibiliography.pdf

[35] Cardoen, B., Demeulemeester, E., Beliën, J.: Operating room planning and scheduling: A literature review. European Journal of Operational Research **201**(3), 921 – 932 (2010)

[36] Carter, M.W.: Diagnosis: Mismanagement of resources. OR/MS Today **29**(2), 26–32 (2002)

[37] Carter, M.W., Blake, J.T.: Using simulation in an acute-care hospital: easier said than done. In: Operations Research and Health Care: A Handbook of Methods and Applications. Kluwer Academic Publishers, New York (2004)

[38] Cattani, K., Schmidt, G.M.: The pooling principle. INFORMS Transactions on Education **5**(2), 47–52 (2005)

[39] Cayirli, T., Veral, E.: Outpatient scheduling in health care: A review of literature. Production and Operations Management **12**(4), 519–549 (2003)

[40] Cayirli, T., Veral, E., Rosen, H.: Designing appointment scheduling systems for ambulatory care services. Health Care Management Science **9**(1), 47–58 (2006)

[41] Ceglowski, R., Churilov, L., Wasserthiel, J.: Combining data mining and discrete event simulation for a value-added view of a hospital emergency department. The Journal of the Operational Research Society **58**(2), 246–254 (2007)

[42] Centeno, M.A., Albacete, C., Terzano, D.O., Carrillo, M., Ogazon, T.: A simulation study of the radiology department at JMH. In: Proceedings of the 2000 Winter Simulation Conference, pp. 1978–1984 (2000)

[43] Center for Healthcare Operations Improvement and Research (CHOIR): ORchestra Bibliography (2011). Retrieved January 15, 2011 from http://www.choir.utwente.nl/en/orchestra

[44] Cochran, J.K., Bharti, A.: A multi-stage stochastic methodology for whole hospital bed planning under peak loading. International Journal of Industrial and Systems Engineering **1**(1), 8–36 (2006)

[45] Cochran, J.K., Bharti, A.: Stochastic bed balancing of an obstetrics hospital. Health Care Management Science **9**(1), 31–45 (2006)

[46] Cohen, J.W.: The single server queue, *North-Holland Series in Applied Mathematics and Mechanics*, vol. 8, second edn. North-Holland Publishing Co., Amsterdam (1982)

[47] Conforti, D., Guerriero, F., Guido, R.: Optimization models for radiotherapy patient scheduling. 4OR: A Quarterly Journal of Operations Research **6**(3), 263–278 (2007)

[48] Costa, A.X., Ridley, S.A., Shahani, A.K., Harper, P.R., De Senna, V., Nielsen, M.S.: Mathematical modelling and simulation for planning critical care capacity. Anaesthesia **58**(4), 320 (2003)

[49] Criswell, M., Hasan, I., Kopach, R., Lambert, S., Lawley M.and McWilliams, D., Trupiano, G., Varadarajan, N.: Emergency department divert avoidance using petri nets. Proceedings of the IEEE International Conference on System of Systems Engineering (2007). 1–6

[50] Currie, C.T., Hoy, D., Tierney, A.J., Bryan-Jones, J., Lapsley, I.: HipMod: Development of a multi-agent audit-based computer simulation of hip fracture care. Health Informatics Journal **9**(3), 183 (2003)

[51] Cutler, D.M.: Empirical evidence on hospital delivery under prospective payment (1990). MIT

[52] Dafny, L.S.: How do hospitals respond to price changes? The American Economic Review **95**(5), 1525–1547 (2005)

[53] Dansky, K.H., Miles, J.: Patient satisfaction with ambulatory healthcare services: waiting times and filling time. Hospital and health services administration **77**, 42–165 (1997)

[54] De Bruin, A.M., Koole, G.M., Visser, M.C.: Bottleneck analysis of emergency cardiac in-patient flow in a university setting: an application of queueing theory. Clinical & Investigative Medicine **28**(6), 316–7 (2005)

[55] Dean, B., van Ackere, A., Gallivan, S., Barber, N.: When should pharmacists visit their wards? An application of simulationto planning hospital pharmacy services. Health Care Management Science **2**(1), 35–42 (1999)

[56] Dean, B.S., Allan, E.L., Barber, N.D., Barker, K.N.: Comparison of medication errors in an American and a British hospital. American Journal of Health-System Pharmacy **52**(22), 2543–2549 (1995)

[57] Derlet, R.W., Richards, J.R.: Overcrowding in the nation's emergency departments: Complex causes and disturbing effects. Annals of emergency medicine **35**(1), 63–68 (2000)

[58] Dexter, F.: Bibliography of operating room management articles. Retrieved October 10, 2008 from http://www.franklindexter.net (2009)

[59] Dexter, F., Blake, J.T., Penning, D.H., Sloan, B., Chung, P., Lubarsky, D.: Use of linear programming to estimate impact of changes in a hospital's operating room time allocation on perioperative variable costs. Anesthesiology **96**(3), 718–724 (2002)

[60] Dexter, F., Lubarsky, D.A.: Using length of stay data from a hospital to evaluate whether limiting elective surgery at the hospital is an inappropriate decision. Journal of Clinical Anesthesia **16**(6), 421–425 (2004)

[61] Dexter, F., Macario, A., Traub, R.D.: Enterprise-wide patient scheduling information systems to coordinate surgical clinic and operating room scheduling can impair operating room efficiency. Anesthesia & Analgesia **91**(3), 617–626 (2000)

[62] van Dijk, N.M.: On hybrid combination of queueing and simulation. In: Proceedings of the 2000 Winter simulation Conference, pp. 147–150 (2000)

[63] van Dijk, N.M., van der Sluis, E.: To pool or not to pool in call centers. Production and Operations Management **17**(3), 296–305 (2004)

[64] van Dijk, N.M., van der Sluis, E.: Practical optimization by OR and simulation. Simulation Modelling Practice and Theory **16**(8), 1113 – 1122 (2008)

[65] van Dijk, N.M., van der Sluis, E.: Pooling is not the answer. European Journal of Operational Research **197**(1), 415–421 (2009)

[66] Dimakou, S., Parkin, D., Devlin, N., Appleby, J.: Identifying the impact of government targets on waiting times in the NHS. Health Care Management Science **12**(1), 1–10 (2009)

[67] Drummond, A.J.: No room at the inn: overcrowding in Ontario's emergency departments. Canadian Journal of Emergency Medicine **4**(2), 91–7 (2002)

[68] Dudgeo, N.: Canada - Netherlands seminar on health care. Tech. rep., Canadian College of Health Service Executives (2007)

[69] Epstein, R.H., Dexter, F.: Economic analysis of linking operating room scheduling and hospital material management information systems for just-in-time inventory control. Anesthesia & Analgesia **91**(2), 337–343 (2000)

[70] Erlenkotter, D.: Sequencing expansion projects. Operations Research **21**(2), 542–553 (1973)

[71] Everett, J.E.: A decision support simulation model for the management of an elective surgery waiting system. Health Care Management Science **5**(2), 89–95 (2002)

[72] Fackrell, M.: Modelling healthcare systems with phase-type distributions. Health Care Management Science **12**(1), 11–26 (2009)

[73] Federgruen, A., Groenevelt, H.: M/G/c queueing systems with multiple customer classes: characterization and control of achievable performance under nonpreemptive priority rules. Management Science **34**(9), 1121–1138 (1988)

[74] Fletcher, A., Worthington, D.J.: What is a 'generic' hospital model? (2007). EPrint: Retrieved October 13, 2008, from http://eprints.lancs.ac.uk/7051/1/004583.pdf

[75] Fletcher, A., Worthington, D.J.: What is a generic hospital model? A comparison of generic and specific hospital models of emergency patient flows. Health Care Management Science **12**(4), 374–391 (2009)

[76] Folmer, K., Mot, E.: Diagnosis and treatment combinations in Dutch hospitals. Tech. rep., CPB Netherlands Bureau for Economic Policy Analysis (2003)

[77] Freidenfelds, J.: Capacity expansion-analysis of simple models with applications. North Holland (1981)

[78] Gallivan, S., Utley, M., Treasure, T., Valencia, O.: Booked inpatient admissions and hospital capacity: mathematical modelling study. BMJ **324**(7332), 280–282 (2002). DOI 10.1136/bmj.324.7332.280

[79] Gorunescu, F., McClean, S.I., Millard, P.H.: A queueing model for bed-occupancy management and planning of hospitals. Journal of the Operational Research Society **53**, 19–24 (2002)

[80] Green, L.V., Savin, S.: Reducing delays for medical appointments: A queueing approach. Operations Research **56**(6), 1526–1538 (2008)

[81] Green, L.V., Savin, S., Murray, M.: Providing timely access to care: What is the right patient panel size? Joint Commission Journal on Quality and Patient Safety **33**(4), 211–218 (April 2007)

[82] Griffiths, J.D., Price-Lloyd, N., Smithies, M., Williams, J.E.: Modelling the requirement for supplementary nurses in an intensive care unit. Journal of the Operational Research Society **56**(2), 126–133 (2005)

[83] Gross, D.: Fundamentals of queueing theory. Wiley-India (2008)

[84] Gross, M.: Wait times: the appropriateness of the methodology and how they affect patients. Canadian Journal of Surgery **47**(3), 167–169 (2004)

[85] Guinet, A., Chaabane, S.: Operating theatre planning. International Journal of Production Economics **85**(1), 69–81 (2003)

[86] Hall, R.W.: Patient Flow: Reducing Delay in Healthcare Delivery, first edn. Springer, New York (2006)

[87] Hanratty, B., Robinson, M.: Coping with winter bed crises. New surveillance systems might help. British Medical Journal **319**(7224), 1511–1512 (1999)

[88] Hans, E.W., van Houdenhoven, M., Hulshof, P.J.H.: A framework for health care planning and control (2011). EPrint: Retrieved April 27, 2011, from http://eprints.eemcs.utwente.nl/19571/

[89] Harper, P.R.: A framework for operational modelling of hospital resources. Health Care Management Science **5**(3), 165–173 (2002)

[90] Harper, P.R., Shahani, A.K.: Modelling for the planning and management of bed capacities in hospitals. Journal of the Operational Research Society **53**(1), 11–18 (2002)

[91] Harper, P.R., Shahani, A.K., Gallagher, J.E., Bowie, C.: Planning health services with explicit geographical considerations: a stochastic location-allocation approach. Omega **33**(2), 141 – 152 (2005)

[92] Harris, R.A.: Hospital bed requirements planning. European Journal of Operational Research **25**(1), 121–126 (1986)

[93] Hermes Consulting Solutions: Panel sizer software (2011). Retrieved January 15, 2011 from http://www.panelsizer.com

[94] Hochang, L.: Project selection problems for production-inventory-distribution scheduling in manufacturing plants. In: Computer-Aided Design, Engineering, and Manufacturing, pp. 1–27. CRC Press (2000)

[95] Hopp, W.J., Spearman, M.L.: Factory physics: foundations of manufacturing management. McGraw-Hill, Boston (2001)

[96] van Houdenhoven, M., Hans, E., Klein, J., Wullink, G., Kazemier, G.: A norm utilisation for scarce hospital resources: Evidence from operating rooms in a Dutch university hospital. Journal of Medical Systems **31**(4), 231–236 (2007)

[97] van Houdenhoven, M., van Oostrum, J.M., Wullink, G., Hans, E., Hurink, J.L., Bakker., J., Kazemier, G.: 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)

[98] Hreinsson, E.B.: Hydroelectric project sequencing using heuristic techniques and dynamic programming. In: Power Systems Computation Conference, Cascais, Portugal (1987)

[99] Huang, X.M.: Patient attitude towards waiting in an outpatient clinic and its applications. Health Services Management Research **7**(1), 2–8 (1994)

[100] Huang, X.M.: A planning model for requirement of emergency beds. Mathematical Medicine and Biology **12**(3-4), 345–353 (1995). DOI 10.1093/imammb/12.3-4.345

[101] Huckman, R.S., Zinner, D.E.: Does focus improve operational performance? Lessons from the management of clinical trials. Strategic Management Journal **29**(2), 173–193 (2008)

[102] Hyer, N.L., Wemmerlöv, U., Morris, J.: Performance analysis of a focused hospital unit: The case of an integrated trauma center. Journal of Operations Management **27**(3), 203–219 (2009)

[103] Institute for Healthcare Inprovement (IHI): Optimizing patient flow: Moving patients smoothly through acute

care settings (2003). Retrieved June 23, 2008 from http://www.ihi.org/IHI/Results/WhitePapers

[104] Irwin, R.S., Rippe, J.M.: Irwin and Rippe's intensive care medicine. Wolters Kluwer Health/Lippincott Williams & Wilkins (2008)

[105] Janssen, A., van Leeuwaarden, J., Zwart, B.: Corrected asymptotics for a multi-server queue in the Halfin-Whitt regime. Queueing Systems **58**(4), 261–301 (2008)

[106] Jebali, A., Hadj Alouane, A.B., Ladet, P.: Operating rooms scheduling. International Journal of Production Economics **99**(1-2), 52–62 (2006)

[107] Jiang, L., Giachetti, R.: A queueing network model to analyze the impact of parallelization of care on patient cycle time. Health Care Management Science **11**(3), 248–261 (2008)

[108] Johansson, L.: Decentralisation from acute to home care settings in Sweden. Health policy **41**(1), 131–143 (1997)

[109] Joustra, P., van der Sluis, E., van Dijk, N.M.: To pool or not to pool in hospitals: a theoretical and practical comparison for a radiotherapy outpatient department. Annals of Operations Research **178**(1), 77–89 (2010)

[110] Jun, J.B., Jacobson, S.H., Swisher, J.R.: Application of discrete-event simulation in health care clinics: A survey. Journal of the Operational Research Society **50**(2), 109–123 (1999)

[111] van de Ketterij, J.L., Schaepkens, F.F.J.M., de Vries, P.G.: DBC 2003 what's in it for me. Health information developments in the Netherlands **1**(5) (2002)

[112] Kim, S.C., Horowitz, I.: Scheduling hospital services: the efficacy of elective-surgery quotas. Omega **30**(5), 335–346 (2002)

[113] Kim, S.C., Horowitz, I., Young, K.K., Buckley, T.A.: Analysis of capacity management of the intensive care unit in a hospital. European Journal of Operational Research **115**(1), 36–46 (1999)

[114] Kim, S.C., Horowitz, I., Young, K.K., Buckley, T.A.: Flexible bed allocation and performance in the intensive care unit. Journal of Operations Management **18**(4), 427–43 (2000)

[115] Koizumi, N., Kuno, E., Smith, T.E.: Modeling patient flows using a queuing network with blocking. Health Care Management Science **8**(1), 49–60 (2005)

[116] Kokangul, A.: 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)

[117] Kolisch, R., Hartmann, S.: Experimental investigation of heuristics for resource-constrained project scheduling: An update. European Journal of Operational Research **174**(1), 23 – 37 (2006)

[118] Kolker, A.: Process modeling of ICU patient flow: Effect of daily load leveling of elective surgeries on ICU diversion. Journal of Medical Systems **33**(1), 1–14 (2008)

[119] Kommer, G.J.: A waiting list model for residential care for the mentally disabled in the Netherlands. Health Care Management Science **5**(4), 285–290 (2002)

[120] Kotiadis, K.: Extracting a conceptual model for a complex integrated system in healthcare. In: J. Garnett, S. Brailsford, S. Robinson, S. Taylor (eds.) Proceedings of the Third Operational Research Society Simulation Workshop, pp. 235–245. The Operational Research Society (2006)

[121] Kremitske, D.L., West, D.J.: Patient-focused primary care: a model. Hospital Topics **75**(4), 22 – 28 (1997)

[122] Kwak, N.K., Kuzdrall, P.J., Schmitz, H.H.: The GPSS simulation of scheduling policies for surgical patients. Management Science **22**(9), 982–989 (1976)

[123] Lane, D.C., Monefeldt, C., Rosenhead, J.V.: 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)

[124] Langabeer, J.R., Ozcan, Y.A.: The economics of cancer care: longitudinal changes in provider efficiency. Health Care Management Science **12**(2), 192–200 (2009)

[125] Leonard, K.J., Rauner, M.S., Schaffhauser-Linzatti, M.M., Yap, R.: The effect of funding policy on day of week admissions and discharges in hospitals: the cases of Austria and Canada. Health Policy **63**(3), 239 – 257 (2003)

[126] Leung, G.M.: Hospitals must become Focused Factories. BMJ: British Medical Journal **320**(7239), 942–943 (2000)

[127] Litvak, E., Long, M.C.: Cost and quality under managed care: Irreconcilable differences. American Journal of Managed Care **6**(3), 305–312 (2000)

[128] Litvak, N., van Rijsbergen, M., Boucherie, R.J., van Houdenhoven, M.: Managing the overflow of intensive care patients. European Journal of Operational Research **185**(3), 998–1010 (2008)

[129] Lovejoy, W.S., Li, Y.: Hospital operating room capacity expansion. Management Science **48**(11), 1369–1387 (2002)

[130] Lowery, J.C., Martin, J.B.: Evaluation of an advance surgical scheduling system. Journal of Medical Systems **13**(1), 11–23 (1989)

[131] Luss, H.: Operations research and capacity expansion problems: A survey. Operations Research **30**(5), 907–947 (1982)

[132] Mahapatra, S., Koelling, C.P., Patvivatsiri, L., Fraticelli, B., Eitel, D., Grove, L.: Pairing emergency severity index 5-level triage data with computer aided system design to improve emergency department access and throughput. In: Proceedings of the 2003 Winter Simulation Conference, vol. 2, pp. 1917–1925 (2003)

[133] Mandelbaum, A., Reiman, M.I.: On pooling in queueing networks. Management Science **44**(7), 971–981 (1998)

[134] Martin, S., Smith, P.C.: Rationing by waiting lists: an empirical investigation. Journal of Public Economics **71**(1), 141–164 (1999)

[135] Massey, W.A., Whitt, W.: Networks of infinite-server queues with nonstationary Poisson input. Queueing Systems **13**(1), 183–250 (1993)

[136] Masursky, D., Dexter, F., O'Leary, C.E., Applegeet, C., Nussmeier, N.A.: Long-term forecasting of anesthesia workload in operating rooms from changes in a hospital's local population can be inaccurate. Anesthesia & Analgesia **106**(4), 1223–1231 (2008)

[137] Matta, M.E., Patterson, S.S.: Evaluating multiple performance measures across several dimensions at a multi-facility outpatient center. Health Care Management Science **10**(2), 173–194 (2007)

[138] Mayhew, L.: On the effectiveness of care co-ordination services aimed at preventing hospital admissions and emergency attendances. Health Care Management Science **12**(3), 269–284 (2009)

[139] McDonald, J.H.: Handbook of Biological Statistics, 2nd ed. Sparky House Publishing, Baltimore, Maryland. (2009)

[140] McGowan, J.E., Truwit, J.D., Cipriano, P., Howell, R.E., VanBree, M., Garson, A., Hanks, J.B.: Operating room efficiency and hospital capacity: Factors affecting operating room use during maximum hospital census. Journal of the American College of Surgeons **204**(5), 865–871 (2007)

[141] McLaughlin, C.P., Yang, S., van Dierdonck, R.: Professional service organizations and focus. Management Science **41**(7), 1185–1193 (1995)

[142] McManus, M.L., Long, M.C., Cooper, A., Mandell, J., Berwick, D.M., Pagano, M., Litvak, E.: Variability in surgical caseload and access to intensive care services. Anesthesiology **98**(6), 1491–1496 (2003)

[143] van der Meer, R.B., Rymaszewski, L.A., Findlay, H., Curran, J.: Using OR to support the development of an integrated musculo-skeletal service. Journal of the Operational Research Society **56**(2), 162–172 (2005)

[144] van Merode, G.G., Groothuis, S., Schoenmakers, M., Boersma, H.H.: Simulation studies and the alignment of interests. Health Care Management Science **5**(2), 97–102 (2002)

[145] Ministerie van Infrastructuur en Milieu: Sneller-beter (2011). Retrieved April 15, 2011 from http://www.snellerbeter.nl/ (see [68] for English Summary)

[146] Morin, T.L.: Optimal sequencing of capacity expansion projects. Journal of the Hydraulics Division **99**(9), 1605–1622 (1973)

[147] Morin, T.L.: Multidimensional sequencing rule. Operations Research **23**(3), 576–580 (1975)

[148] Morin, T.L., Shin, Y.S.: Optimal expansion of flood control systems, vol. 1. Northwestern University (1977)

[149] Murray, M., Berwick, D.M.: Advanced access: Reducing waiting and delays in primary care. Journal of the American Medical Association **289**(8), 1035–1040 (2003)

[150] Murray, M., Davies, M., Boushon, B.: Panel size: How many patients can one doctor manage? Family Practice Management **14**(4), 44 (2007)

[151] Neebe, A.W., Rao, M.R.: The discrete-time sequencing expansion problem. Operations Research **31**(3), 546–558 (1983)

[152] Neebe, A.W., Rao, M.R.: Sequencing capacity expansion projects in continuous time. Management Science **32**(11), 1467–1479 (1986)

[153] Newman, K.: Towards a new health care paradigm. Patient-focused care. The case of Kingston Hospital Trust. Journal of Management in Medicine **11**(6), 357–371 (1997)

[154] Nguyen, J.M., Six, P., Antonioli, D., Glemain, P., Potel, G., Lombrail, P., Le Beux, P.: A simple method to optimize hospital beds capacity. International Journal of Medical Informatics **74**(1), 39–49 (2005)

[155] Nguyen, J.M., Six, P., Parisot, R., Antonioli, D., Nicolas, F., Lombrail, P.: A universal method for determining intensive care unit bed requirements. Intensive Care Medicine **29**(5), 849–852 (2003)

[156] O'Kane, P.C.: A simulation model of a diagnostic radiology department. European Journal of Operational Research **6**(1), 38–45 (1981)

[157] van Oostrum, J.M., Bredenhoff, E., Hans, E.W.: Suitability and managerial implications of a master surgical scheduling approach. Annals of Operations Research **178**(1), 91–104 (2010)

[158] van Oostrum, J.M., van Houdenhoven, M., Hurink, J.L., Hans, E.W., Wullink, G., Kazemier, G.: A master surgical scheduling approach for cyclic scheduling in operating room departments. OR Spectrum **30**(2), 355–374 (2008)

[159] Oudhoff, J.P., Timmermans, D.R.M., Rietberg, M., Knol, D.L., Van der Wal, G.: The acceptability of waiting times for elective general surgery and the appropriateness of prioritising patients. BMC Health Services Research **7**(1), 32–44 (2007)

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

[161] Pitt, D.F., Noseworthy, T.W., Guilbert, J., Williams, J.R.: Waiting lists: management, legalities and ethics. Canadian Journal of Surgery **46**(3), 170–175 (2003)

[162] Ramis, F.J., Palma, J.L., Baesler, F.F.: The use of simulation for process improvement at an ambulatory surgery center. In: Proceedings of the 2001 Winter Simulation Conference, pp. 1401–1404 (2001)

[163] Rauner, M.S., Schaffhauser-Linzatti, M.M.: Impact of the new Austrian inpatient payment strategy on hospital behavior: a system-dynamics model. Socio-Economic Planning Sciences **36**(3), 161 – 182 (2002)

[164] Renwick, M., Gillett, S., Liu, Z.: Long-stay older patients in acute hospitals: are they bed blockers? Australian health review **15**(3), 284–98 (1992)

[165] Robbins, M.J., Jacobson, S.H.: Pediatric vaccine procurement policy: The monopsonist's problem. Omega (forthcoming) (2011)

[166] Ross, S.M.: Stochastic Processes. Wiley (1995)

[167] Rotondi, A.J., Brindis, C., Cantees, K.K., DeRiso, B.M., Ilkin, H., Palmer, J.S., Gunnerson, H.B., Watkins, W.D.: Benchmarking the perioperative process. I. Patient routing systems: a method for continual improvement of patient flow and resource utilization. Journal of Clinical Anesthesia **9**(2), 159–69 (1997)

[168] Rubin, S.G., Davies, G.H.: Bed blocking by elderly patients in general-hospital wards. Age and Ageing **4**(3), 142–147 (1975)

[169] Sakasegawa, H.: An approximation formula $l_q \approx \alpha\rho^\beta/(1-\rho)$. Annals of the Institute of Statistical Mathematics **29**(1), 67–75 (1977)

[170] Samaha, S., Armel, W.S., Starks, D.W.: The use of simulation to reduce the length of stay in an emergency department. In: Proceedings of the 2003 Winter Simulation Conference, vol. 2 (2003)

[171] Santibáñez, P., Begen, M., Atkins, D.: Managing surgical waitlists for a British Columbia health authority (2005). Retrieved October 10, 2008, from http://www.chcm.ubc.ca/docs/Managing-Surgical-Waitlists-British-Columbia-Health-Authority.pdf

[172] Santibáñez, P., Begen, M., Atkins, D.: 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)

[173] Schneider, J.E., Miller, T.R., Ohsfeldt, R.L., Morrisey, M.A., Zelner, B.A., Li, P.: The economics of specialty hospitals. Medical Care Research and Review **65**(5), 531 (2008)

[174] Sier, D., Tobin, P., McGurk, C.: Scheduling surgical procedures. Journal of the Operational Research Society **48**(9), 884–891 (1997)

[175] Skinner, W.: Manufacturing: The Formidable Competitive Weapon. John Wiley & Sons Inc, New York (1985)

[176] Sobolev, B., Harel, D., Vasilakis, C., Levy, A.: Using the statecharts paradigm for simulation of patient flow in surgical care. Health Care Management Science **11**(1), 79–86 (2008)

[177] Sokal, S.M., Craft, D.L., Chang, Y., Sandberg, W.S., Berger, D.L.: Maximizing operating room and recovery room capacity in an era of constrained resources. Archives of Surgery **141**(4), 389 (2006)

[178] Sutherland, J., Hamm, J., Hatcher, J.: Adjusting case mix payment amounts for inaccurately reported comorbidity data. Health Care Management Science **13**(1), 65–73 (2010)

[179] Syam, S.S., Ct, M.J.: A location-allocation model for service providers with application to not-for-profit health care organizations. Omega **38**(3-4), 157 – 166 (2010)

[180] Takakuwa, S., Shiozaki, H.: Functional analysis for operating emergency department of a general hospital. In: Proceedings of the 2004 Winter simulation Conference, pp. 2003–2011 (2004)

[181] Taylor, K., Lane, D.: Simulation applied to health services: opportunities for applying the system dynamics approach. Journal of Health Services Research & Policy **3**(4), 226–32 (1998)

[182] Taylor, R.G.: A general form for the capital projects sequencing problem. In: Proceedings of the 21st international conference on computers and industrial engineering, pp. 47–50. Elsevier Science Publishers Ltd., Essex, UK (1997)

[183] Testi, A., Tanfani, E.: Tactical and operational decisions for operating room planning: Efficiency and welfare implications. Health Care Management Science **12**(4), 363–373 (2009)

[184] Testi, A., Tanfani, E., Torre, G.: A three-phase approach for operating theatre schedules. Health Care Management Science **10**(2), 163–172 (2007)

[185] Tijms, H.C.: A First Course in Stochastic Models. John Wiley and Sons, New York (2003)

[186] Tiwari, V., Heese, H.S.: Specialization and competition in healthcare delivery networks. Health Care Management Science **12**(3), 306–324 (2009)

[187] Utley, M., Gallivan, S., Davis, K., Daniel, P., Reeves, P., Worrall, J.: Estimating bed requirements for an intermediate care facility. European Journal of Operational Research **150**(1), 92–100 (2003)

[188] Utley, M., Gallivan, S., Treasure, T., Valencia, O.: 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)

[189] Vanberkel, P.T., Blake, J.T.: A comprehensive simulation for wait time reduction and capacity planning applied in general surgery. Health Care Management Science **10**(4), 373–385 (2007)

[190] Vanberkel, P.T., Boucherie, R.J., Hans, E.W., Hurink, J.L., Lent, W.A.M., Harten, W.H.: Accounting for inpatient wards when developing master surgical schedules. Anesthesia & Analgesia (forthcoming) (2011)

[191] Vanberkel, P.T., Boucherie, R.J., Hans, E.W., Hurink, J.L., Lent, W.A.M., Harten, W.H.: An exact approach for relating recovering surgical patient workload to the master surgical schedule. Journal of the Operational Research Society (forthcoming) (2011)

[192] Vanberkel, P.T., Boucherie, R.J., Hans, E.W., Hurink, J.L., Litvak, N.: Reallocating resources to focused factories: A case study in chemotherapy. In: J.T. Blake, M.W. Carter (eds.) International Perspectives on Operations Research and Health Care: Proceedings of the 34th Meeting of the European Working Group on Operational Research Applied to Health Services, pp. 152–164 (2010)

[193] Vanberkel, P.T., Boucherie, R.J., Hans, E.W., Hurink, J.L., Litvak, N.: A survey of health care models that encompass multiple departments. International Journal of Health Management and Information **1**(1), 37–69 (2010)

[194] Vasilakis, C., Marshall, A.H.: Modelling nationwide hospital length of stay: opening the black box. Journal of the Operational Research Society **56**(7), 862–869 (2005)

[195] Vissers, J.M.H.: Patient flow-based allocation of inpatient resources: A case study. European Journal of Operational Research **105**(2), 356–370 (1998)

[196] Vissers, J.M.H., Beech, R.: Health Operations Management: Patient Flow Logistics in Health Care. Routledge, Oxon (2005)

[197] Wachtel, R.E., Dexter, F.: Tactical increases in operating room block time for capacity planning should not be based on utilization. Anesthesia & Analgesia **106**(1), 215–226 (2008)

[198] Wartman, S.A., Morlock, L.L., Malitz, F.E., Palm, E.A.: Patient understanding and satisfaction as predictors of compliance. Medical Care **21**(9), 886–91 (1983)

[199] Westert, G.P., Burgers, J.S., Verkleij, H.: The Netherlands: regulated competition behind the dykes? BMJ **339**(3397), 839–842 (2009)

[200] Whitt, W.: Approximating a point process by a renewal process, I: Two basic methods. Operations Research **30**(1), 125–147 (1982)

[201] Whitt, W.: The queueing network analyzer. Bell System Technical Journal **62**(9), 2779–2815 (1983)

[202] Whitt, W.: Partitioning customers into service groups. Management Science **45**(11), 1579–1592 (1999)

[203] Wickramasinghe, N., Bloemendal, J.W., De Bruin, A.K., Krabbendam, J.J.: Enabling innovative healthcare delivery through the use of the focused factory model: the case of the spine clinic of the future. International Journal of Innovation and Learning **2**(1), 90–110 (2005)

[204] Wiecek, M.M., Ehrgott, M., Fadel, G., Figueira, J.R.: Multiple criteria decision making for engineering. Omega **36**(3), 337 – 339 (2008). Special Issue on Multiple Criteria Decision Making for Engineering

[205] Winston, W.L.: Operations research: applications and algorithms, third edn. International Thomson Publishing, Tampa (1994)

[206] Wolstenholme, E.: A patient flow perspective of UK health services: exploring the case for new "intermediate care" initiatives. System Dynamics Review **15**(3), 253–271 (1999)

[207] Wong, C., Geiger, G., Derman, Y.D., Busby, C.R., Carter, M.W.: Redesigning the medication ordering, dispensing, and administration processin an acute care academic health sciences centre. In: Proceedings of the 2003 Winter Simulation Conference (2003)

[208] Worthington, D.J.: Hospital waiting list management models. Journal of the Operational Research Society **42**(10), 833–843 (1991)

[209] Wright, M.B.: The application of a surgical bed simulation model. European Journal of Operational Research **32**(1), 26–32 (1987)

[210] Zijm, W.H.M.: Towards intelligent manufacturing planning and control systems. OR Spectrum **22**(3), 313–345 (2000)

[211] Zonderland, M.E., Boer, F., Boucherie, R.J., de Roode, A., van Kleef, J.: Redesign of a university hospital preanesthesia evaluation clinic using a queuing theory approach. Anesthesia & Analgesia **109**(5), 1612–1621 (2009)

List of abbreviations

ASA approximate solution approach
BCCA British Columbia Cancer Agency
CDU chemotherapy day unit
CHOIR Center for Healthcare Operations Improvement and Research
DI diagnostic imaging
DRG diagnosis related group
ED emergency department
EOF economies of focus
EOS economies of scale
GP general practitioner
ICU intensive care unit
IID independent and identically distributed
ILP integer linear program
LAB laboratory medicine
LOS length of stay
MSS master surgical schedule
NCI Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital
OR operating room
PACU post-anesthesia care unit
PSP project sequencing problem

Summary

In this thesis we address a number of challenging problems related to health care logistics. These problems are motivated by hospital managers who collaborated in the research, and the results are applied at their hospitals. The general results and solution approaches presented in this thesis are also valid in other hospital settings.

To position the research we review quantitative health care literature to examine the extent to which models encompass multiple hospital departments and account for department-to-department interactions. We provide a general overview of the relationships which exist between major hospital departments and describe how these relationships are accounted for by researchers. Our review of literature found that researchers often confine models to single departments due to system complexity and the uncertain nature of patient flows (Chapter 2). Using and developing techniques from queueing theory, mathematical programming, and simulation, we demonstrate how these characteristics can be coped with by solving multiple strategic, tactical, and operational problems faced by our partner hospitals.

Using queueing theory we model the complex and uncertain relationship between capacity, case mix and patient mix. With parameters provided by this queueing model, we formulate a combinatorial optimization problem to maximize the hospital's remuneration under a fee-for-service financing system. We thus provide a methodology for optimizing strategic capacity and case mix planning decisions. Ex-

act solutions can be found with integer linear program solvers and approximate solutions with dynamic programming (Chapter 3).

A second strategic problem is deciding whether (and to what extent) to pool resources within hospitals. Due to the uncertainty of patient arrivals and the economies of scale found in the pooled departments, access time will typically be worse in unpooled departments. However, if the service time is sufficiently lower in the unpooled departments, due to more focused care, the opposite is true. Using queueing theory we derive general results stating the extent to which focused care must decrease service time in an unpooled department in order to compensate for the lack of economies of scale. The main characteristics influencing economies of scale losses are clinic load, proportional size of the patient groups, resource divisions and appointment length variability (Chapter 4).

At a tactical level, physicians and hospital managers must decide how many patients a single physician can effectively be accountable for (i.e. panel size). We formalize an extension to existing models allowing the panel size to be a random variable which accounts for the uncertainty in patient flows. Using queueing theory we provide general results related to capacity planning and provide strategies for reaching and maintaining a panel size that meets certain performance criteria (Chapter 5).

Developing a surgical schedule that does not overwhelm inpatient wards is a complex problem, given that surgery durations and patient length of stays are uncertain. Using applied probability, we develop a solution approach for the tactical level master surgical scheduling problem. Our approach, used to develop a new master surgical schedule at the collaborating hospital, is readily repeatable and has been used at multiple Dutch hospitals. Using a case study, and by comparing predicted ward occupancies with post-implementation ward occupancies, we validated the approach (Chapter 6).

An operational level problem faced by many pharmacies is deciding when to prepare medication. This problem is complex because medica-

tions are expensive and have a limited shelf life, and uncertain, because patient no-shows are common in hospitals. Analyzing this problem for a chemotherapy pharmacy, our case study predicted waiting times could be decreased by 30 minutes while only increasing pharmacy costs by 1-2%. The research led to analytic approximations (validated with discrete event simulation) usefull for predicting patient waiting times and costs in any pharmacy. Our analysis in Chapter 7 led to a new pharmacy policy which has been implemented at the collaborating hospital.

Acknowledgements

During my PhD study I was extremely fortunate to be supervised by four erudite and energetic people. I am thankful to Erwin Hans, first of all, for taking a risk and helping me secure my PhD position and moreover for his seemingly endless supply of enthusiasm which is both infectious and motivating. I'm thankful to Richard Boucherie for his frank assessments of my research, both good and bad, which taught me appreciably and built my confidence as a researcher. I'm thankful to Johann Hurink for teaching me how to write in an organized and coherent manner and for not simply conceding that I should know best since *I'm* the native English speaker. I'm thankful to Nelly Litvak for all the times she made me stop what I was doing in order to force me to return to the fundamentals. I strive for practical results, but Nelly helped me understand that "there is nothing more practical than good theory".

In addition to my supervisors, I wish to thank the many colleagues that I worked with and the many friends that I made. Specifically, I wish to thank:

My colleagues at the Netherlands Cancer Institute for partnering with me and supporting my research: Wim van Harten, Wineke van Lent, Hans Schoo, Arthur Dernison, Tom van der Mijden, Inge Masselink. My colleagues from the University of British Columbia for hosting my research visit: Martin Puterman, Scott Tyldesley, Pablo Santibanez, Vincent Chow, Ruben Aristizabal, Kevin Huang.

My fellow PhD students from OMPL, LogiDOC and SOR for helping me with a variety of things, not the least being: LNMB courses, filing my tax returns and reading my Dutch mail. Being surrounded by such astute and scholarly individuals set the bar very high which surely contributed to the success of my study. Although I won't miss the coffee, I'll certainly miss the coffee breaks.

My Enschede friends: Renze, Maartje, Des, Clare, Tom, Paul, Fede, John, Denis, Katya, Domi, Zhofi, Leendert, Simona, Sabrina, Aimee, Cheryl, Marcel, Erwin, Hana, Elias, Elia, Luisa, Orlando and also the members of the slapping studs hockey team. You certainly helped me make the most of my time here. In this case, I will miss both the good times and the Grolsch.

And of course my family: Albert and Wies for the many relaxing weekends spent sitting around your kitchen table. My parents, siblings and in-laws for tolerating us being so far away, for so long and for your unconditional support. Caitlin for always smiling, even when upset, and for reminding me of what matters. And Connie, for starting (and finishing) this adventure with me and for all the love and support along the way.

Peter
Enschede, May 2011.

About the author

Peter T. Vanberkel, PEng, MASc

Peter was born in Antigonish, Nova Scotia, Canada on September 14th, 1981. He attended St. Francis Xavier University and Dalhousie University, where he earned a bachelor's degree in engineering and a master's of applied science degree. His PhD study at the University of Twente culminates with this thesis.

He is a registered professional engineer with Engineers Nova Scotia and has worked as an industrial engineer at Canadian Tire Corporation (internship), Michelin Tire Canada Ltd. (internship), the Capital District Health Authority and the IWK Health Centre. As a researcher he has worked at the Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, the British Columbia Cancer Agency, the University of British Columbia and the University of Twente.

List of publications

Journal Articles

Vanberkel P.T., Boucherie R.J., Hans E.W., Hurink J.L., van Lent W.A.M., van Harten W.H., Accounting for Inpatient Wards when developing Master Surgical Schedules, Anesthesia & Analgesia (forthcoming)

Vanberkel P.T., Boucherie R.J., Hans E.W., Hurink J.L., van Lent W.A.M., van Harten W.H., An exact approach for relating recovering surgical patient

workload to the master surgical schedule, Journal of the Operational Research Society (forthcoming)

Vanberkel P.T., Boucherie R.J., Hans E.W., Hurink J.L., Litvak N., Efficiency evaluation for pooling resources in health care, OR Spectrum, (forthcoming)

Vanberkel P.T., Boucherie R.J., Hans E.W., Hurink J.L., Litvak N., (2010) A Survey of Health Care Models that Encompass Multiple Departments, International Journal of Health Management and Information, **1** (1) 37 - 69

Vanberkel P.T., Blake J.T., (2007) A Comprehensive Simulation for Wait Time Reduction and Capacity Planning Applied in General Surgery. Health Care Management Science, **10** (4) 373-385, Special Issue on Simulation in Health Care.

Working Papers (under review)

Vanberkel P.T., Boucherie R.J., Hans E.W., Hurink J.L., Optimizing the Strategic Patient Mix

Masselink I.H.J., van der Mijden T.L.C., Litvak N., Vanberkel P.T., Preparation of chemotherapy drugs: planning policy for reduced waiting times.

Hulshof P.J.H., Vanberkel P.T., Boucherie R.J., Hans E.W., van Houdenhoven M., van Ommeren J.C.W., An Analytical Comparison of the Patient-to-Doctor Policy and the Doctor-to-Patient Policy in the Outpatient Clinic

van Lent W.A.M., Vanberkel P.T., van Harten W.H., A review on the relation between simulation and improvement

Refereed Conference Proceedings

Vanberkel P.T., Boucherie R.J., Hans E.W., Hurink J.L., Litvak N., van Lent W.A.M., van Harten W.H., (2009) Reallocating Resources to Focused Factories: A Case Study in Chemotherapy, In Proceedings of the European Working Group on Operational Research Applied to Health Services (ORAHS). Edited by J.T. Blake and M.W. Carter.

Blake J.T., Vanberkel P.T., Dunbar M., Malloy L., Hennigar A, Storey M., (2004) Wait List Management at Capital District Health Authority. In Proceedings of the 30th Meeting of the European Working Group on Operational Research Applied to Health Services (ORAHS). Edited by M. Lagergren.

Dunbar M., Blake J.T., Vanberkel P.T., Molloy L., Hennigar A., (2008) Development of a Wait List Computer Simulation Model For Elective Orthopaedic Surgery. Journal of Bone Joint Surgery Br **9** 0-B: 111-a

Campbell M., Vanberkel P.T., (2007) Just-in-time patient scheduling in an Eye Care Clinic. In 37th Annual Conference Proceedings of the Atlantic Schools of Business. Edited by J. Blake.

Book Contributions

Vanberkel P.T., Hans E.W., (2009) Holistic healthcare modeling. A viewpoint on managing the complete patient care chain. Contribution to the book: Operational Research Applied to Health Services in Action (ISBN 978-83-7493-409-1)

Vanberkel P.T., (2009) Population Density and Location of Future Information Societies. Contribution to the book: Living the ICT Future (ISBN 978-90-365-2963-1)

Vanberkel P.T., Blake J.T., (2008) Quantitative Modelling for Wait Time Reduction, (ISBN 978-3-693-09370-4)

Conference Proceedings / Technical Papers (Not Refereed)

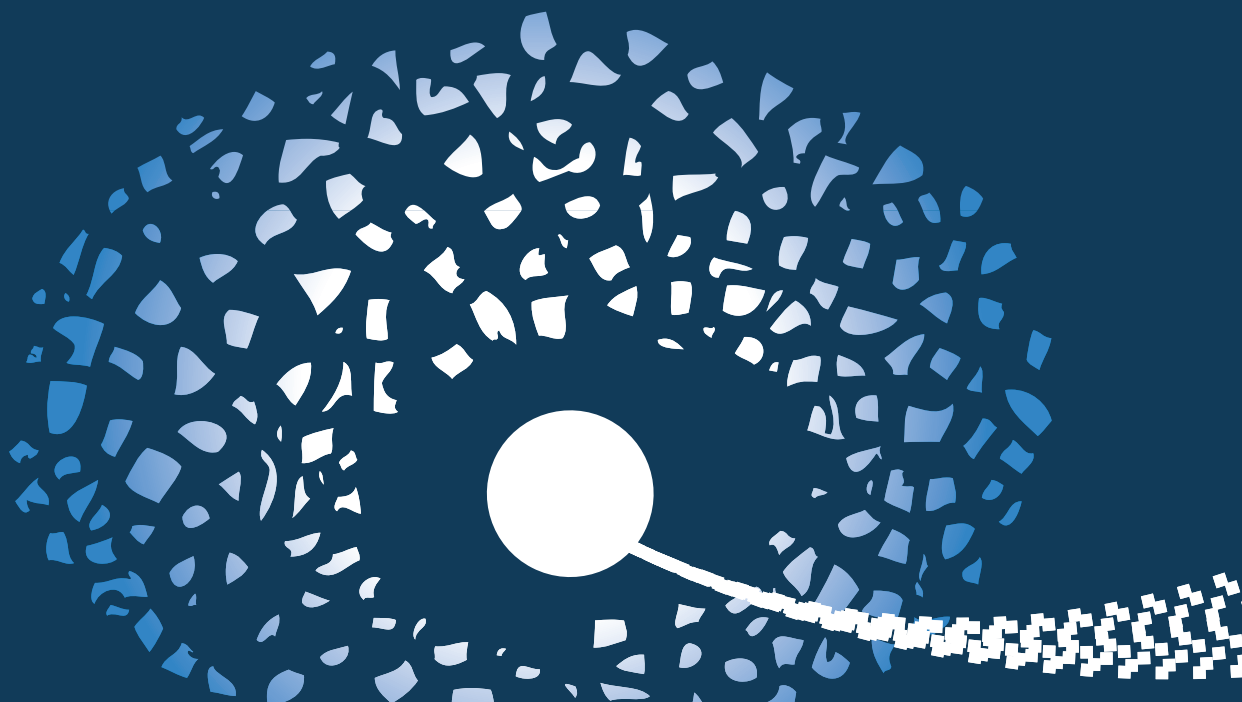
Vanberkel P.T., Boucherie R.J., Hans E.W., Hurink J.L., Litvak N., (2009) Designing for Economies of Scale vs. Economies of Focus in Hospital Departments. In Proceedings of Innovation in Design and Management of Healthcare Facilities and Healthy Environments Research Day, Rotterdam, The Netherlands

Vanberkel P.T., (2004) Predicting the Results of Resource Changes: A Computer Simulation in a Health Care Environment. Winning Paper, Devonyx Technical Paper Competition, Canadian Society of Industrial Engineers Conference

Thesis

Vanberkel P.T., (2006) The Impact of the Quantity and Use of Resources on Elective Wait Times: A Comprehensive Simulation Applied in General Surgery. Master's Thesis, Department of Industrial Engineering, Dalhousie University

In this thesis a number of challenging problems related to health care logistics are addressed. These problems are motivated by hospital managers who collaborated in the research, and the results are applied at their hospitals. The general results are valid in other hospital settings and the solution approaches used to cope with system complexity and patient flow uncertainty are novel. Using and developing techniques from queueing theory, mathematical programming, and simulation, multiple strategic, tactical and operational problems are solved, demonstrating how complexity and uncertainty can be coped with in health care settings.



Beta
Research School for Operations Management and Logistics

UNIVERSITY OF TWENTE.
Department of Applied Mathematics
Operational Methods for Production and Logistics
CTIT Dissertation Series No. 11-198