A Computational Approach to Patient Flow Logistics in Hospitals

A catalogue record is available from the Eindhoven University of Technology Library.

Hutzschenreuter, Anke Kristine

A Computational Approach to Patient Flow Logistics in Hospitals/ by Anke Kristine Hutzschenreuter.
- Eindhoven: Technische Universiteit Eindhoven, 2010. - Proefschrift. -

ISBN 978-90-8891-165-1

NUR 982

Keywords: Decision Support Systems / Health care logistics / Computational Intelligence / Multi-agent simulation / Online Multi-objective Optimization

The work in this thesis has been carried out under the auspices of Beta Research School for Operations Management and Logistics.

Beta Dissertation Series D131

Printed by Proefschriftmaken.nl
Cover design by Marijke Timmermans

A Computational Approach to Patient Flow Logistics in Hospitals

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
rector magnificus, prof.dr.ir. C.J. van Duijn, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op woensdag 26 mei 2010 om 16.00 uur

door

Anke Kristine Hutzschenreuter

geboren te Ulm, Duitsland

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr.ir. J.A. La Poutré
en
prof.dr.ir. J.W.M. Bertrand

Copromotor:
dr. P.A.N. Bosman

To my father

Contents

Chapter 1

Introduction

Today, the European health care systems are facing great pressure. One reason for this is the increased demand for care due to a rapidly ageing population. At the same time new medical technologies keep emerging that improve diagnosis and treatment possibilities but also increase costs for health care provision. The health care expenditures in the Netherlands, for example, have increased from 8 percent of the gross national product in 2000 to 13 percent in 2005[1]. Therefore, the efficient organization of the health care system has become a major political issue. In this context, hospitals are of particular interest as they yield the single largest costs in the health care system. In the Netherlands hospital costs amount to approximately thirty percent of the total health care expenditures[2]. In order to reduce health care expenditures, many European countries, like the Netherlands, have introduced a free market health care system to increase the competition among care providers. Moreover, in 2005 the Dutch government introduced a case-based reimbursement system (diagnosis related groups) for part of the hospital services that reward more efficient utilization of resources. Furthermore, patients increasingly include factors such as reputation, patient service and waiting times in their choice of health care service provider. Due to the increased cost pressure, competition and patients' consumer awareness, many hospitals face the need to optimize their processes in favor of cost optimization and reduced patient waiting times. In order to decrease costs hospitals need to increase the utilization of their resources and reduce the patients' duration of admission. Increasing the resource utilization, however,

[1]Data obtained from the European World Health Organization (WHO), http://data.euro.who.int/

[2]Obtained from Statistics Netherlands (Dutch: Centraal Bureau voor de Statistiek) for the year 2005, http://www.cbs.nl/

may lead to bottlenecks that cause blocking of patient flow and consequently increasing patient waiting times. Thus, the efficient planning of care services in hospitals becomes increasingly important.

Admission control plays an important role in the efficient planning of care provision [1, 99]. Admission control is concerned with selecting a mix of patients to be admitted to the hospital as inpatients such that the available resource capacity and the demand for health care services are matched. Through the combination of the different care requirements of different patient types the available resources can be used in a more efficient and effective way. As stated in Groot [37] goals of admission control are amongst others a high utilization of the available capacity, smooth patient flow resulting in minimal length of stay of the patients and improved patient service.

Closely related to admission control is hospital resource management that targets the efficient deployment of resources, for example operating rooms or beds, when and where they are needed. Allocating resources to the different units in the hospital affects the patient mix that can be admitted to the hospital. Clearly, by in- or decreasing the capacity at a hospital unit the flow of patients at the respective unit is in- or decreased. But as patient pathways often involve more than one hospital unit and possibly share resources with other pathways, an allocation decision may also (indirectly) influence other patient flows and thus the possible patient admissions. For example, if resources at a unit that is involved in one patient flow are increased, this may compromise the flow of other patients at shared resources that may then be mainly occupied by the first increased patient flow. In a straightforward way admission decisions influence the resources that are needed at the different units in the hospital. Therefore, resource management and admission control are important and coupled managerial issues to be considered in order to improve hospital operations.

In many hospitals the planning of patient admissions and the allocation of hospital resources are major managerial issues, especially due to the complex relationship between resources, utilization and patient throughput for different patient groups [40]. One reason for this is the uncertainty that is inherent in hospital operations. First, patient arrivals are stochastic. Emergency patients arrive in urgent need for care and require immediate admission to the hospital. Also, the arrival of elective patients is uncertain, however, their arrival may be buffered by a waiting list. Second, the treatment processes of patients are often stochastic. Complications may occur that require a patient's transfer to another care unit than anticipated and also the duration of treatment of a patient at a care unit is stochastic. Moreover, the planning task is highly complex, as hospital planners need to

consider multiple patient treatment processes that typically involve several hospital units. Often, resources (e.g. at the Intensive Care unit) are shared by multiple treatment processes. Thus, hospital resource management and patient admission planning are complex and highly dynamic problems.

In this thesis we develop planning techniques for decision support on patient admission control and hospital resource management. We consider multiple hospital care units, multiple patient groups with stochastic treatment processes and uncertain resource availability due to the overlapping patient treatment processes. In the remainder of this chapter, we present a general description and model of the hospital domain and patient flows in Section 1.1, the planning problems at hand and the aim of our research are described in Section 1.2. Then, we outline the approach taken in this work and define the scope of our work in Section 1.3. We end this chapter with an overview of related work on hospital planning in Section 1.4, an outline and roadmap for the remainder of this thesis in Section 1.5 and an overview of the publications this thesis is based upon in Section 1.6.

1.1 Description and model of problem domain

In this section a description of the hospital domain and the patient flows is given. First, we describe the hospital domain with the organizational structure of a hospital and the different hospital units relevant for this study. Then, we provide a description of patient flows and treatment processes. Finally, we present a generic domain model including patient flows and constraints.

1.1.1 Characteristics of the hospital domain

In general, a hospital can be divided into several, medically specialized, care units [26, 59, 81]. Hospital care units like nursing wards provide treatment and monitoring and are typically dedicated to a medical specialty such as orthopedics or cardiothoracic surgery. In the terminology of industrial organizational theory, the term workstation would be used to denote the different departments in a hospital as the responsibility for the patients' treatment processes remains at the respective specialists who has admitted the patient to the hospital. However, since the term 'hospital (care) unit' is commonly used in the field, we will adopt this nomenclature in the remainder of this thesis.

Often, hospital care units are shared by different specialties. Examples of shared care units are the operating room (OR) unit, where medical

specialties are assigned time slots for performing surgical procedures, and the intensive care unit (ICU), where patients with serious to life-threatening diseases are monitored. Often, the ICU is divided into several subunits characterized by different care levels. Care levels indicate the intensity of care and monitoring. Within the ICU, we distinguish three care levels: intensive care (IC), high care (HC) and medium care (MC), in decreasing order. Another important part of the ICU is the post anesthesia care unit (PACU) where patients recovering from anesthesia are monitored. Unless complications occur, patients stay at the PACU only for a few hours before returning home or to another hospital unit. Some hospitals also have designated ICU areas for medical specialties, e.g. the Coronary Care Unit (CCU) for heart disease. A description of the care provided at the different care units and their admission criteria is given in Table 1.1.

Unit	**Description**	**Patient admission criteria**
Intensive Care (IC)	Highly technical care and monitoring including artificial respiration or internal monitoring	Patients with serious to life-threatening health condition
Cardiac Care Unit (CCU)	Comparable to IC with extensive heart monitoring and testing equipment, staff specialized in heart conditions and procedures	Patients after a heart attack or major cardiac surgery.
High Care (HC)	Highly technical care and monitoring	Postoperative patients
Medium Care (MC)	Care focussed on revalidation, less technical than IC and HC, but intensive alertness and additional facilities, e.g. ECG or oxygen saturation monitoring; "step-down" after IC/HC	Postoperative and/or trauma patients that do not satisfy admission criteria for IC/HC, but are too care-intensive for regular ward
Post anesthesia care unit (PACU)	Highly technical postsurgery monitoring and care	Postoperative patients recovering from anesthesia
Ward	Revalidation and care linked to corresponding specialty	Patients of corresponding specialty

Table 1.1: Hospital care units with description of provided care and patient admission indication

For providing patient care at a hospital unit, resources are required. Relevant resources are ORs and hospital beds. The availability of facil-

ity resources may be temporary, for example, ORs are typically available between 8 a.m. and 5 p.m. Hospital beds may also be opened only for a predetermined time period which is typically the case at the PACU.

In order to accommodate patients at the appropriate care level, back-up capacity may be used. This means that an additional bed is opened at the respective care unit or that a patient is temporarily accommodated at another unit until a regular bed is available. For example, the CCU may serve as back-up for the ICU. Usage of back-up capacity is undesired by hospital management as it may introduce additional operational effort for the medical staff involved in patient and/or bed transports. Also, back-up capacity usage may affect other patient groups that require the resources at the back-up unit.

Schedules of shared resources, like ORs, are managed locally by the different units. Typically, each unit applies its own (medical) priorities and preferences that may be based on medical guidelines, working habits, etc. and are specific to the medical domain considered. Moreover, patient admissions and transfers are planned in a decentralized fashion. Information concerning patient admissions and transfers is solely communicated to other hospital units if a patient needs to be transferred to the respective facility. Thus, planning in hospitals has strong decentralized features.

1.1.2 Patient flows in a hospital

Patient flows can be classified based on how hospital resources are needed. This classification results in an outpatient flow (patients who visit the hospital, clinic, or associated facility for diagnosis or treatment but are not hospitalized) and inpatient flow (patients who are admitted to the hospital and stay overnight or for an indeterminate time, usually several days or weeks). In this thesis we focus on the latter patient flow.

In the following, we distinguish between elective surgical patients and emergency patients in urgent need for intensive care. We assume that surgical patients are always put on a waiting list[3]. Surgical and emergency inpatient flows typically involve multiple care units, such as the specialities' wards, the OR and postoperative care departments. The patients' postoperative care requirement is often uncertain, depending on the complexity of the surgical procedures. The treatment processes considered in this thesis

[3]This assumption also allows for the treatment of surgical emergency patients such that emergency patients are given highest priority and are either placed on top of the waiting list or admitted instead of an elective patient. However, this is not included in the analysis.

are complex, i.e. they are characterized by uncertainty and involve multiple care units.

Furthermore, we assume that patients can be grouped on the basis of their resource consumption. The resource consumption is determined by the required treatment steps, the involved resources and the respective duration, see e.g. [1, 99]. The objective for patient grouping is to identify and anticipate the resource need of different patient groups [63]. In the medical domain, various grouping and classification techniques are developed and used [33]. For example, diagnosis related groups (DRGs) [33] provide an aggregated way of patient grouping. A machine learning approach for patient grouping based on detailed process data is presented in [66]. Alternatively, techniques like knowledge elicitation from medical specialists or statistical data analysis could be used to determine hospital-specific patient grouping.

In this thesis, we focus on cardiothoracic inpatient flows and their interaction with other surgical and emergency patient flows. The cardiothoracic patient group was chosen because the associated patient flows are in general well-defined, which is advantageous as patient flow mining is beyond the scope of this thesis. Moreover, cardiothoracic patients represent a large patient population. The cardiothoracic treatment process involves the surgical treatment of coronary heart and lung diseases, e.g (open-)heart surgery for coronary artery bypass grafting and heart valve replacement. Due to the complex surgical procedures it is not very predictable which postoperative care the patients require. The treatment involves several care units, i.e. the OR, ICU and a corresponding ward. The OR and ICU are typically shared with other surgical and emergency patients. The interaction can have a great impact on the patient flows because it results in limited and uncertain resource availability at these units. As the resources at the OR, the ICU and the wards are not used by outpatient flows, we do not include these flows in the analysis and model.

The models that are used in this thesis for the hospital domain and patient pathways are further elaborated in the following section.

1.1.3 Domain and patient flow model

The underlying model in this thesis is comprised of two principal components: a network of specialized hospital units and patient pathways according to which patients of different groups (cf. Section 1.1.2) are flowing through the network.

Domain model & resource allocation

The hospital units provide treatment and monitoring for (parts of) the patients' treatment process for which resources are required. Hospital resources comprise, for example, diagnostic facilities (such as CT scanners), equipment (such as heart rate monitors), specialized staff, operating rooms (ORs) and hospital beds [87]. In this thesis, we restrict our focus to ORs and hospital beds, as they are crucial for hospital production and their utilization is of major managerial importance. Here, we consider personnel and equipment scheduling as subsequent problem to OR and bed allocation, where facility allocations serve as constraints to personnel and equipment scheduling. For the latter problems sophisticated techniques have been proposed in the literature, e.g. [3, 21, 22, 54, 68, 92], which may be combined with our approach. In the following we assume that resources are fully staffed and equipped with specialized facilities.

A resource allocation specifies how resources are assigned to the different units. Here, we consider the temporary assignment of resources. The assignment of ORs, for example, is typically done in half-day OR sessions. Moreover, allocated hospital beds may available only for certain time periods. For instance, beds at the PACU are open for a limited period of time during and after OR working hours. This factor is incorporated in our model as the temporary availability of (post)surgical resources may affect the workload at the ICU and other care units. Furthermore, the allocation of resources may change over time, for example the bed capacity at nursing wards during the weekend may be reduced.

The allocation of resources is associated with costs for required staff, materials, etc. that need to be taken into account by the hospital management in efficiency considerations.

Patient flow model

We define a patient pathway (in the following also referred to as patient path) of a patient group as the sequence of required treatment operations and their respective duration. In the hospital context, the treatment duration is typically referred to as Length of Stay (LoS). The LoS is modeled as a random variable that follows a predefined probability distribution. As discussed in Marazzi et al. [63], typical models for patients' LoS are asymmetric distributions with outliers towards high values of LoS (right-skewed). Widely used models are Lognormal, Gamma and Weibull distributions [63].

The patient pathways considered in this thesis are complex and stochas-

tic, i.e. pathways involve multiple hospital units and the routing between adjacent treatment steps is stochastic. The stochastic routing reflects the possibility of complications that require a patient transfer to another unit than expected, for example the ICU, and is modeled by conditional probability distributions.

An example of a surgical pathway is depicted in Figure 1.1. The patient pathway model is illustrated using a graph-like structure where the involved hospital units are represented by nodes that are connected by arrows that depict the patient flows. Here, the pathway comprises the surgery and the postoperative care of a group of surgical patients. Postoperative complications may occur with a probability, $Pr(HC|OR)$, and require the admission to the HC prior to revalidation at the ward. Without complications, with probability $1-Pr(HC|OR)$, patients return to the ward directly after surgery in this example.

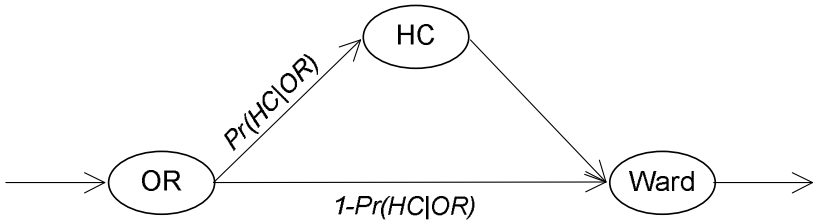


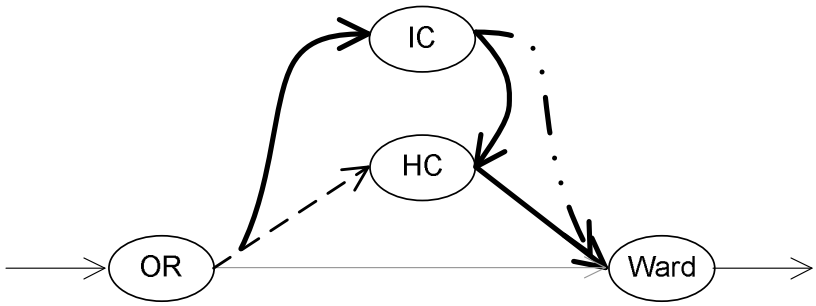
Figure 1.1: Example for a surgical patient pathway including postoperative complications that require transfer to the HC

Patient pathways indicate the resource need of patients in a hospital. The actual flow of patients through the network of care units, however, is also determined by the availability of resources. This means that a patient may be admitted or transferred to a unit that is not indicated by the patient's pathway if no bed is available at the destined unit. The possibilities for adapting a patient's pathway are

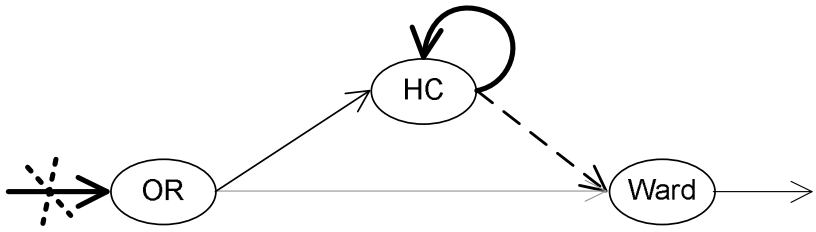
1. the patient is (temporarily) admitted to another care unit,
2. the concerned patient remains (temporarily) admitted to the current unit.

The first possibility is restricted to units of equal or higher care level in order to ensure the quality of the patients' care. The latter option may require the usage of back-up capacity. Both options comprise the possibility of a later patient transfer to the originally indicated unit if a bed becomes available. The two possibilities are depicted in Figure 1.2 for the example

in Figure 1.1. Here, the dashed arrows represent that infeasible patient transfers to the indicated units due to resource unavailability. The bold arrows depict the possible adaptations of the patient pathways.



(a) temporary admission to another care unit



(b) temporarily continued admission

Figure 1.2: Examples for temporarily adapting the patient pathway depicted in Figure 1.1 through (a) admitting the patient temporarily to the IC due to unavailability of HC bed and (b) continued stay at HC in case of unavailability of ward bed after complications (resource unavailability is represented by dashed arrows)

The transfer probabilities, determined by $Pr(HC|OR)$, in Figure 1.1 thus refer to the general occurrence of complications and their type. The adaptation of the patient pathway, however, is determined based on the resource availability at the different units by mutual agreement between the corresponding care units' planners and physician(s) in charge. Moreover, adapting the pathways induces discrete and disruptive behavior in the system.

Using back-up capacity means that an additional bed is opened at the care unit for a short period of time. An additional bed can be accomplished through a (temporarily) increased workforce or shifting a bed from another hospital unit. Therefore, back-up capacity usage is undesired by hospital staff and managers and should be accounted for in assessing a resource allocation.

In practice, the decisions whether a patient is transferred to an alternative unit, which units are considered for alternative transfer and whether patients are retransferred to the originally indicated unit depend on the availability, the demand and urgency for the corresponding hospital beds and the transfer policy employed by the hospital unit(s).

In general, we assume that the amount of patient flow into a care unit equals the amount of flow out of the care unit, shifted by a non-constant time per patient. Without this assumption patients could possibly remain at a hospital care unit for an infinite period of time.

The flow of patients into the hospital is determined by the admission scheme (also referred to as admission schedule), the demand for care and the resource availability. The admission scheme specifies the (maximal) number of patients per patient group to be admitted to the hospital on a certain day and thus determines the planned patient mix in the hospital. Admission schemes are typically set up and handled in a decentralized fashion by the different care units and specialties due to the organizational structure (cf. Section 1.1.1). In practice, the time horizon for an admission scheme varies per specialty and hospital between one day and several weeks or months. The actual patient admissions are typically limited by the current demand for care, like for example for emergency patients whose arrival is uncertain, and the resource availability at the concerned care units. Thus, an admission scheme specifies an upper bound for the actual number of patients to be admitted. The higher the number of possible patient admissions in the admission scheme, the greater is the impact of the patient demand and resource availability on the actual patient admissions. The situation where solely the patient demand and resource availability determines the actual patient admissions is in the remainder also referred to as unconstrained admission control, cf. Chapter 3.

If the number of patients to be admitted exceeds the number of available beds, a multitude of clinical variables determines which patients are admitted. In our model, we represent this medical choice by a stochastic process that randomly selects patients for available beds (excluding back-up capacity). The same decision-making model is used for patient transfers between care units. As our model is set up in a generic way, incorporating a more elaborate model for the clinical decision-making into our model would be straightforward. However, the medical decision-making in patient care is beyond the scope of this thesis and is therefore not taken into consideration.

The flow of patients leaving the hospital, i.e. after completion of the patients' treatment processes, can have different destinations. For example, patients can be discharged to their homes or other care facilities, like for

instance rehabilitation centers or elderly homes. Patient discharges to other care facilities may be restricted by local admission schemes or bed availability. This constraint is not considered in our model since the focus of our work is on the processes and resources within a hospital. Depending on the diagnosis or illness, also mortality is a possible discharge destination. In our model, the different discharge possibilities are considered on an aggregated level.

1.2 Problem description and contributions

1.2.1 Problem definition

As described in Section 1.1, the problem domain considered in this thesis is characterized by autonomously planning and deciding hospital care units and stochastic and heterogeneous patient flows. In such a complex system, the unbalanced utilization of hospital resources is a major problem. Unbalanced resource utilization means that periods with under-utilized resource capacity are alternated by periods with resource scarcity, e.g. at the ICU. To a certain extent unbalanced utilization cannot be avoided in a stochastic environment. However, an inappropriate resource allocation and improper planning of patient admissions can potentially worsen the situation. This is due to the time-dependency of resource allocation and patient admission decisions, meaning that decisions taken now may cause an unnecessarily unbalanced resource utilization in the current and future periods and thereby affect possible patient admissions in the future.

Unbalanced resource utilization has two implications. On the one hand, allocated resources are used inefficiently. Moreover, costs may incur for unused resource capacity. The costs for unused OR capacity, for example, may amount to several thousand euro per hour for an empty OR. On the other hand, the scarcity of resources causes major planning difficulties and patient flow disturbances. First, the scarcity of resources may lead to the unintended accommodation of patients at units that are not indicated by their corresponding patient pathway, cf. Section 1.1.3. This may lead to capacity conflicts and patients being structurally admitted to non-indicated units which possibly requires unintended patient transfers and transports in addition to the disturbance of the patient flows. Second, resource scarcity may lead to the the blocking of patient flows which leads to cancelations of patient admissions and possibly the costly cancelation of surgeries. This, in turn, may compromise the patients' satisfaction. If back-up capacity may be used as described in Section 1.1.3 this may alleviate cancelations of

admissions and surgeries for one specialty but possibly cause the disturbance of other patient flows that are not included in the model.

Moreover, the decentralization of hospital organizations complicates the centralized management control in hospitals. Specialties may mistrust improved admission schedules proposed by a decision support system that does not take the decentralized way of decision-making into account. Consequently, hospital specialties may not adhere to proposed admission schemes and thereby introduce unforeseen additional occupancy fluctuation in the network of hospital units.

1.2.2 Research goal

Given the problems described in Section 1.2.1, the research goal of the work presented in this thesis can be summarized as follows:

> *Develop methods and techniques for decision support for hospital patient flow logistics taking into account the high degree of uncertainty, heterogeneity and decentralization present in the hospital domain, to facilitate an efficient usage of hospital resources.*

Based on the research goal this thesis addresses the following questions:

- How can we design a fine-grained simulation that reflects the decentralized decision-making in the hospital domain and that is based on real-life case study incorporating (medical) guidelines and business rules? For this aim, the simulation should incorporate models for complex patient treatment processes involving multiple hospital units and stochastic resource need.

- How can we predict future hospital resource usage given the current resource occupancy? To answer this question we need to determine which (de-)centralized information the prediction should be based upon. Moreover, we need to develop a model to realistically capture future fluctuating resource usage. For computational efficiency a prediction function for this problem is desirable.

- How can we optimize the number of allocated hospital resources such that the resource allocations at multiple hospital care units are coordinated? For this purpose we need to identify relevant goals to be taken into consideration for the optimization. Also, we need to determine an appropriate and efficient solution method for this optimization problem.

- How can we design adaptive policies for allocating resource to the different hospital units that dynamically respond to changes in the hospital environment and optimize such allocation policies with respect to multiple conflicting objectives? Also, future resource occupancy should be anticipated and taken into account when designing and optimizing adaptive resource allocation policies.

1.2.3 Contributions

The main contributions made by answering the research questions in Section 1.2.2 are:

- An agent-based simulation for coordinating cardiothoracic, other surgical and emergency patient flows has been developed where the involved specialties and hospital units are each represented by an agent. The simulation has been designed based on knowledge elicitation in the from of interviews with domain experts during a case study and statistical data analysis. The simulation has been extensively validated by simulation experiments and the simulation and results has been approved by domain experts and planners from the case study hospital (Chapter 2).

- Two methods for predicting future resource usage given current bed occupancy and planned patient admissions have been developed that can assist in proactive decision-making: forward simulation using the agent-based simulation and supervised learning using artificial neural networks. We have assessed the precision of the developed techniques and demonstrated that the obtained predictions improve benchmark forecasts derived from hospital practice (Chapter 3).

- An approach for the optimization of the resource allocation at multiple hospital units has been developed using an evolutionary algorithm. The algorithm determines optimal resource allocations according to multiple conflicting criteria simultaneously. This approach has been shown using computational experiments to improve the current hospital practice for resource allocation (Chapter 4).

- Policies for adaptive resource allocation have been designed that can anticipate future resource usage and that are implementable and understandable for planners in hospital practice. A policy optimization approach for dynamic multi-objective optimization has been developed to determine the policy parameters using an evolutionary algorithm.

Computational experiments show that the optimized resource allocations optimized in Chapter 4 can be considerably improved using our adaptive allocation policies. Moreover, the policies can be further improved by including prediction information (Chapter 5).

1.3 Approach

The focus of this thesis is on the computational aspects of health care management science. Thus, our research is set in-between the fields of computer science and management studies. Specifically, our approach combines techniques from agent-based simulation and computational intelligence for decision support in health care which are designed and evaluated on the basis of a complex realistic hospital setting. The keystones of our approach are described below.

1.3.1 Case study

For a realistic study of patient planning in a hospital setting, the techniques presented in this thesis were developed in cooperation with the Catharina Hospital Eindhoven (CHE), the Netherlands. The CHE is a large university-affiliated general hospital that offers international state-of-the-art medicine for, amongst others, cardiothoracic surgery (CTS) and intensive care in addition to the required basic medical care. An extensive case study was performed at the CTS department and the ICU which comprised several interviews with medical specialists, planners and managers and an extensive statistical analysis of data from the hospital information system. On the basis of the CHE case study we studied the requirements for a simulation in this setting. The simulation is described in detail in Chapter 2 and incorporates multiple planned surgical and emergency patient flows. The parameters of the patient flows in the simulation were determined based on the CHE patient flows and the initial resource allocation was set according to the CHE situation. Moreover, the decision-making at the different care units at the CHE inspired the policies employed by the units in the simulation and their parameter values. Using the CHE parameter settings as simulation instance, we validated the simulation and showed that the results obtained from the simulation compare well to the outcomes realized by the human planners at the CHE. Furthermore, the CHE parameter settings were used for the experimental evaluation of the prediction approaches for future resource occupancy and the resource management optimization presented in the different chapters of this thesis. Our results obtained under

the realistic settings used in our evaluations demonstrate the applicability of the developed techniques in a real-world problem setting. Furthermore, our results should also be implementable in other hospital settings with comparable problem features in terms of heterogeneity, stochasticity and decentralization. This is due to the fact that we used the CHE case study for our requirements analysis for this type of problem. Furthermore, the CHE case study setting is sufficiently generic for surgical (particularly with regard to CTS) and emergency patient flows. Moreover, we varied the CHE parameters in our evaluations, showing the robustness of the solutions.

1.3.2 Agent-based simulation

As discussed in Section 1.1, hospitals often show a distributed organizational structure. They are divided into several autonomous hospital units, each associated with a medical specialty. Patient admission decisions are taken locally and indicated patient transfers are negotiated among the concerned care units that each apply their own decision criteria, e.g. bed reservation policies, (medical) priorities or preferences. These domain features are reminiscent of multiagent systems (MAS), making such systems natural candidates for modeling the problem domain and supporting decision rules that are developed in this thesis.

Although there is no generally agreed definition of software agents in the computer science research community, a commonly used definition is given in Weiss [101]:

> *An agent is a computer module that is situated in some environment, and that is capable of autonomous action in this environment in order to meet its design goals.*

A MAS is a system that is composed of multiple agents that communicate and interact with each other in order to solve one or more tasks. In the domain at hand, the agents are models for the real-life entities in the hospital. Specifically, we designed software agents in a MAS that realistically model hospital care units with individual decision-making policies, as in the CHE case, in a detailed fashion and used them in an agent-based simulation (ABS). Thus, an ABS approach closely matches hospitals' organizational structure and allows a detailed modeling of actual decision-making characteristics.

The agents' task is the coordination of patient flow through patient admissions, surgery scheduling and patient transfers to the required care units

such that the hospital resources are used efficiently. The agents are situated in a hospital environment that is characterized by stochastic changes, i.e. complications that require unexpected patient transfers, emergency patient arrivals and uncertain treatment durations. Furthermore, the hospital environment is dynamic, i.e. the environment changes over time due to the patients that are admitted and their demand for care.

In an ABS the interaction and decisions in a MAS are simulated. The simulation is used to assess the effect of the emergent behavior of the agents on the system as a whole. In our approach, an ABS is a refinement of discrete event simulation, where the agent paradigm is used to represent the different hospital entities by autonomous agents in the simulation software.

1.3.3 Computational intelligence

The techniques presented in this thesis combine an agent-based modeling and simulation approach with intelligent planning and scheduling approaches. Specifically, we focused on computational intelligence (CI) techniques for determining scheduling policies. This focus was chosen as exact methods are too computationally intensive and very hard to use in complex and dynamic systems, whereas CI techniques have been shown to be powerful in this problem settings, e.g. in [14, 20, 68, 100].

CI is a branch of the artificial intelligence (AI) research field that considers algorithms such as neural networks, evolutionary algorithms and heuristic search. Algorithms in CI combine elements of learning, adaptation and evolution to be able to handle dynamically changing environments and complex optimization.

Specifically, in this thesis we apply artificial neural networks (ANNs) and heuristic search in Chapter 3 and evolutionary algorithms (EAs) in Chapter 4 and Chapter 5 for the prediction of future resource occupancy and the optimization of resource allocation in hospitals, respectively. The applied CI techniques will be described in detail in the corresponding chapters. Below, a brief introduction of the relevant techniques will be provided. The interested reader is also referred to Bishop [7], Russell and Norvig [82] for an in-depth description of the employed CI techniques.

ANNs provide a general way to define parameterized non-linear functions, inspired by the way in which neurons are connected in the brain. ANNs are commonly used to perform function approximation by fitting the parameters and structure of the network to data, i.e. machine learning. Through their adaptivity, ANNs have been shown to be able to be powerful real-world problem solvers [7, 51, 100].

An EA is a population-based metaheuristic optimization algorithm inspired by biological evolution. Throughout the optimization a set or population of candidate solutions is maintained. Iteratively the candidate solutions are evaluated using a fitness function, selected based on their fitness and new candidate solutions are generated by mutation and recombination operators. EAs have been shown to be very powerful for stochastic optimization, especially in domains where multiple conflicting objectives need to be taken into account [14, 20, 32].

Similar to EAs, heuristic search is an informed search technique where problem-specific knowledge is used to search for an "optimal" solution among a number of candidate solutions [82]. Iterative improvement algorithms often provide a practical approach where an initial solution is selected randomly which is iteratively improved by applying small changes to the current solution. One distinguishes between hill-climbing search that moves to solutions of increasing value and simulated annealing that allow temporarily deteriorating changes. In our work the first iterative improvement technique was applied.

The combination of an agent-based simulation with CI optimization techniques appears to be a promising approach for decision support in a hospital setting where the planning is often performed in a decentralized way. Moreover, it allows for designing and evaluating improved (adaptive) policies, which can then be implemented easily in real life.

1.4 Literature positioning

There is a number of different research areas that are related to the work presented in this thesis. The relevant areas are the fields of operations management, operations research and artificial intelligence, in particular computational intelligence and agent technology. In the following, we provide an outline and give some exemplary references of the work in the different areas and their relevance for the work presented in this thesis. In the respective chapters of this thesis, we will additionally provide a more detailed review of the relevant literature.

1.4.1 Operations Management

In the operations management literature several frameworks can be found that describe the way activities and resources should be managed in a hospital.

Fetter and Freeman [33] propose a framework for product line management [85] based on diagnosis related groups (DRGs). Their control system uses the concept of matrix management in hospitals which means that hospital management is organized both hierarchically at the different care, ancillary units and specialties and laterally across departments for DRG product line management. The proposed financial accounting system distinguishes between clinical and administrative management.

The concept of DRGs is also used for the operations planning and control system presented in Roth and Van Dierdonck [81]. Their framework describes the major components of a hospital resource planning system, i.e. demand forecasting, admission control and capacity planning modules, whose input is translated into resource requirements using the general materials requirements planning logic [85].

A framework for capacity management in health care is proposed in Smith-Daniels et al. [87] where the authors distinguish between decisions concerning the acquisition and the allocation of facility and workforce resources.

A framework for surgical process scheduling is proposed in Blake and Carter [8] that subdivides the problem domain into advance (referring to booking patient admissions in advance), allocation (concerning OR surgery scheduling, cf. Section 1.4.2) and external resource scheduling (relating to booking of required pre- and postoperative care) with a subdivision into strategic, operational and organizational impact of the scheduling decisions.

Vissers et al. [98] present a framework for hospital production control that is based on the framework for general production control in [6]. Due to the framework's focus on production control and its generality, we will use it to position our work and approaches in the related literature. In the following, a brief outline of the framework is given.

The hospital production control framework distinguishes five levels of control. The highest level of *strategic planning* is concerned with the global direction of the hospital in the future, e.g. the extension or addition of a specialism. The second level of control is the *patient volumes planning and control* which involves decisions regarding the required resource capacity and agreements with health insurance companies concerning the patient volumes per diagnosis family. The *resources planning and control* level constitutes the third level of the framework. Here, target resource utilization is defined and the resource usage of the different patient groups and specialties is determined. The fourth control level is called *patient group planning and control* and is concerned with defining treatment policies and urgency criteria and allocating resources to the patient groups. The fifth level, *patient*

planning and control, is concerned with the coupling of resources to single patients and is thus on the operational level of the treatment processes of the individual patients. The work on patient flow control, prediction and hospital resource management presented in this thesis can be positioned on the *patient group planning and control* level within the scope of this framework.

The purpose of the aforementioned frameworks is to describe *what* decisions should be taken, whereas in our work we focus on *how* management decisions should be taken. In this thesis, we present computational models and methods to support hospital management decisions in the presence of complex stochastic patient pathways with overlapping resource requirements.

1.4.2 Operations Research

A range of health care logistics problems has been addressed in the operations research literature, comprising outpatient, e.g. [44, 52, 57] as well as inpatient settings. With respect to inpatient optimization problems the operations research literature has mainly focused on single specialties and care units within a hospital. A significant number of publications has been attended to the problems of OR surgery scheduling, hospital resource management and patient admission scheduling for which examples of related work will be briefly reviewed below[4].

OR surgery scheduling A simulation and a scheduling heuristic to minimize idle time at the OR are presented in Charnetski [23] taking the resulting costs for surgeon, OR staff and facility idle time into account. In Strum et al. [88] a minimal cost analysis is presented for blocks of OR time to be scheduled. The mixed-integer linear optimization approach described in Sier et al. [84] addresses the scheduling of elective surgeries considering slack between procedures and clinical and organizational constraints. Guinet and Chaabane [38] develop a heuristic for OR scheduling that minimizes OR overtime subject to patients' release and due date constraints. Focussing on orthopedic trauma surgery, Bowers and Mould [18] present a simulation and an approximating mathematical model which are used to examine scheduling and resource reservation policies to balance elective and emergency surgeries. The approach presented in Hans et al. [39] addresses the optimization of a master surgery schedule using scheduling heuristics and a local search

[4] Also, the problem of nurse or shift scheduling has been addressed in the operations research literature, e.g. [21, 22, 72, 92]. However, we refrain from discussing this area of research as it lies beyond the focus of this thesis.

approach with focus on the robustness of the resulting schedule regarding possible delays.

In the work presented in this thesis, the issue of OR planning plays a only secondary role as we focus on the coordination of multiple (surgical) patient flows that involve multiple pre- and postoperative care units. OR planning is addressed on a high level and involves the allocation of OR time slots to the CTS patients considering the order of surgeries and cancelations due to unavailable postoperative care resources on an aggregated level. Specifically, OR planning is performed using a heuristic that is based on the a-priori indication for postoperative treatment of CTS and other surgical patients including the corresponding resources' availability.

Hospital resource management Work in this area has primarily addressed general bed allocation decisions. For example, in Vissers [97] aggregated resource allocations are determined by a stepwise approach that is based on long-term projections on future patient flows and resource demand. In Harper and Shahani [40] aggregated allocation policies are evaluated using simulation. The work reported in Kusters and Groot [60] and Vissers et al. [99] provide theoretical results for bed utilization levels. In VanBerkel [93] surgical patient flows are simulated to aid resource allocation decisions at the OR and dedicated care facilities. The work reported in Stummer et al. [89] focusses on determining the location and size of hospital departments in a network of hospitals in a certain region.

In this context, the intensive care unit (ICU) has received special research attention. Ridge et al. [80] present a simulation model based on a case study that is used for the optimization of the ICU bed allocation and propose a reservation policy for emergency patients. In Kim et al. [55] the issue of pooling beds for different specialties at the ICU is addressed. Also, geriatric wards have been addressed specifically, for example [35] where a queueing model is presented.

Moreover, statistical prediction methods have been presented in the literature, e.g. Tandberg and Qualls [91] where a time series approach is developed to forecast patient arrivals and their LoS, or Littig and Isken [61] where a logistic regression model is presented for short-term aggregated occupancy prediction using clinical information.

In our work we consider resource allocations of OR time slots and different types of beds on the level of individual hospital care units taking the stochastic treatment processes and multiple performance measures into account. This combination of problem features has not been addressed in

earlier work but reflects the complexity of real-life hospital resource management. Moreover, we present a novel optimization approach that facilitates the dynamic allocation of resource in hospital practice. In our approach, we use adaptive allocation policies, i.e. parameterized functions, that determine the adaptive resource decisions in different situations given real-time information concerning the resource occupancy. Moreover, we develop detailed prediction methods for forecasting the distribution of future resource occupancy. We use the predicted resource occupancy information in the allocation policies which has not yet been addressed by other authors.

Admission control This problem is addressed in e.g. [1, 9, 37, 58, 99] on different levels of hospital planning, cf. Section 1.4.1. The work in Blake and Carter [9] addresses the optimization of the overall patient case mix on a strategic planning level using a linear programming approach to (1) generate a given required profit and (2) deviate minimally from a predetermined patient mix. On the level of patient group planning and control, Kolesar [58] derives a stationary Markov chain model that calculates the long-run bed occupancy resulting from patient admissions, discharges and transfers. In Groot [37] different admission policies are evaluated in a setting where patient pathways involve the OR and a general pool of postoperative beds. The pathways are characterized by deterministic treatment durations and stochastic arrivals. The admission planning approach in Vissers et al. [99] and its extension in Adan et al. [1] takes the (expected) amount of resources required at the OR and postoperative care units into account.

In our work on predicting future resource occupancy, we consider a heuristic approach to determine admission schemes to control patient arrivals. The complex features of real-life patient treatment processes that we incorporate, i.e. comprising multiple care units in combination with stochastic routing and treatment durations, have not been considered in the approaches mentioned above. In the heuristic approach, we incorporate online resource occupancy information in the decision-making. However, the approach is restricted to local search within the scope of occupancy predictions to incorporate changing patient admissions.

In general, the operations research approaches described above have been shown to be very effective in solving well-defined centralized, aggregated or steady-state optimization problems. However, the techniques have so far found little application in hospital practice, in great part due to the inherent

decentralization of hospital organizations. Due to this, a lower level of modeling and aggregation is needed to consider the inherent diversity in patient attributes and scheduling goals at the different units involved in complex patient flows. In our work, we incorporate both the decentralization and the dynamics of patient flow scheduling using an agent-based simulation approach in combination with computational intelligence optimization techniques. Moreover, our approach considers stochastic patient pathways and their possible interdependency due to overlapping resource requirements of patient flows. Furthermore, in contrast to analytical models our approach is very flexible and can be easily adapted to other settings.

1.4.3 Artificial Intelligence

Related work in the field of artificial intelligence (AI) typically addresses the dynamic nature of hospital logistics and makes use of available medical information systems. An overview of AI planning and scheduling approaches is given in Sypyropoulos [90]. Similarly to the operations research literature, planning and scheduling studies in the AI field have focussed on single units. Also, decentralized decision support approaches have been advocated [69], amongst others for patient scheduling[5]. A literature overview of these two areas is provided below.

Single hospital unit problems Podgorelec and Kokol [79] present a genetic algorithm to tackle the problem of scheduling therapy appointments for multiple types of patients in an outpatient clinic. Vermeulen et al. [95] propose an adaptive mechanism for online adjustments of resource calenders to schedule multiple types of patients with different priorities in a radiology clinic. These approaches are not applicable in the inpatient hospital setting considered in this thesis and do not consider multiple hospital care units that need to coordinate the different patient flows.

Also, several prediction approaches for forecasting patients' LoS and treatment processes have been reported in earlier work. In this area machine learning techniques and ANNS have been applied. For example, Izenberg

[5]In addition to decentralized patient scheduling, several successful MAS approaches have been presented in the areas of modeling and retrieval of medical information/knowledge, clinical decision support for patient monitoring and diagnosis reasoning and documentation of medical treatment activities [67]. Also, the problems of (decentralized) team and shift scheduling has been addressed in earlier work, e.g. [2, 4, 5, 19, 24, 54, 68]. However, these research areas will be omitted in the literature review below as the scope of our work is on patient treatment flow control.

et al. [51] propose an ANN for predicting the mortality of patients after trauma. In Maruster et al. [66] machine learning techniques are presented for grouping patients based on their logistic requirements. In Lowell and Davis [62] ANNs are applied to predict the LoS of a DRG. Walczak et al. [100] use ANNs to predict the exact LoS for different patient groups based on patient characteristics and clinical information. Yeong et al. [102] present an ANN approach for predicting LoS categories using radiological information. In contrast to previous work, we assume that patient treatment process information is available, albeit in a stochastic form. We present different methods for predicting the distribution of future resource occupancy at different hospital care units given the current resource occupancy and planned patient admissions, which has not been addressed in earlier work. Moveover, we apply the obtained predictions in hospital resource management.

Decentralized hospital simulation and scheduling In previous work the effect of decentralized decision-making structures in hospitals has been investigated [59]. Different organizational settings were considered ranging from existing hospital structure with autonomous care and ancillary units over partly decentralized units to fully centralized hospital organizations. In our approach, we model and simulate the existing fully-decentralized hospital structure present in the case study hospital and develop planning methods for providing decision-support.

Earlier work on agent-based hospital simulation has been presented in e.g. [43, 83]. In Sibbel and Urban [83] modeling and design issues for developing general agent-based simulation systems for the hospital domain are discussed. The presented approach is based on a normative reference model for modeling human decision-making behavior by agents. Herrler and Puppe [43] present an agent-based simulation kit based on the simulation environment *SeSAm* [42] to evaluate different scheduling heuristics. Also, two case studies for task scheduling are presented to evaluate the simulation. The work in [43] is embedded in the *Agent.Hospital* framework [56] which also includes an agent-based medical ontology, as well as several normative approaches for processes simulation and patient and staff scheduling in hospitals. In contrast to the approaches described above, our agent-based simulation was built in a bottom-up approach. In our case study at the CHE we analyzed the requirements for a realistic simulation in this setting. As a result, our agent-based simulation is tailored towards inpatient treatment processes and the involved patient flow logistics. As explained in Section 1.1, the associated problem dynamics differ considerably from out-

patient settings. Therefore, earlier agent-based simulation approaches are not applicable in this problem domain. The agents' scheduling behavior in our simulation is modeled by heuristics that are inspired by current hospital scheduling practice. In our work we combine these realistic scheduling heuristics with computational intelligence techniques. In contrast to the simulation approaches described above, our approach allows us to address the dynamic aspects and develop a fast simulation with negligible overhead.

Also, patient scheduling has been addressed in several agent-based approaches. Specifically, task scheduling has been subject of several agent-based planning approaches in the past. In Decker and Li [26] the issue of conflict handling is studied for scheduling patient tests at different ancillary units. The coordination concept is based on on Generalized Partial Global Planning [27, 28] and involves bidding in auctions for time slots taking into account the patients' assigned medical priority and constraints imposed by already scheduled tasks. In a comparable setting, [75, 76, 77] propose an agent-based scheduling system where the planning of patient tests is based on a contract-net protocol [101] and medical wellness functions of patients. In Marinagi et al. [64] the problem of planning and (re-)scheduling patient tests in hospital laboratories is addressed. The proposed approach uses decomposition techniques to divide complex tests into a set of activities and temporal constraints concerning activities and resources. Oddi and Cesta [71] consider the problem of scheduling medical treatments on resources based on clinical protocols for the treatment of patients. The approach uses constraint-based scheduling techniques and a mixed-initiative problem solving mechanism where a constructive solution is further improved by a user or a tabu search algorithm. In Vermeulen et al. [94] a coordination mechanism is presented for exchanging appointment slots in order to improve the patient's schedules in an outpatient setting.

Our work on patient flow logistics is focussed on the process management for surgical and emergency patient flows. In this problem setting, the task scheduling approaches developed in earlier research as described above cannot be applied in a straightforward way. This is due the uncertain availability of resources due to the stochastic patient pathways which imposes additional non-deterministic constraints on the scheduling problem that need to be taken into account. Moreover, the above approaches for patient scheduling have been shown to be efficient in settings with predefined (and partly deterministic) treatment path. This assumption does not hold in the hospital inpatient setting considered in this thesis with patient pathways characterized by stochastic treatment durations and routing. These stochastic factors perturb the resource schedules and thus further complicate the

problem.

1.5 Outline and roadmap of the thesis

The remainder of this thesis is organized as follows.

Chapter 2 describes and validates the agent-based simulation developed in the CTS case study at the CHE. In this chapter we formalize the domain and patient flow model. Moreover, we describe the case study performed at the CHE with the corresponding patient pathways and resource allocation settings. The agent-based simulation is presented with the agents' decision-making policies that were inspired by the case study which also provided the corresponding policy parameters. Several basic and what-if scenarios are presented that demonstrate the functionality of the simulation and applicability for the problem domain.

Chapter 3 focuses on the prediction of future hospital resource occupancy for the model described in Chapter 2. Given the current and planned patient admissions the bed occupancy over a period of several days is predicted. To account for fluctuations during the day, the resource occupancy is modeled by a probability distribution. We present two prediction approaches: forward simulation and supervised learning. Forward simulation is used to forecast the future bed occupancy by estimates of the empirical distribution function calculated based on samples obtained from several simulation runs. For the supervised learning we use (artificial) neural networks. For this approach, the empirical probability distributions of bed occupancy obtained from forward simulation experiments are approximated by Gaussian mixture distributions, i.e. the convex sum of normal distributions, whose parameters are learned by the neural network. We evaluate the sample size needed to obtain accurate and precise predictions using forward simulation and show the feasibility of the supervised learning approach. The forward simulation prediction approach will be used further throughout the thesis.

Chapter 4 is concerned with the optimization of hospital resource management in a network of care units. We present a multi-objective evolutionary optimization approach that uses the simulation presented in Chapter 2 as grey-box evaluation. The three conflicting objectives used in the optimization are: maximal patient throughput, minimal

resource costs and minimal usage of back-up capacity. The optimized resource allocations improve benchmarks obtained from hospital practice in resource management. Moreover, we determine the algorithmic settings required to obtain accurate optimization results at reasonable computational costs.

Chapter 5 addresses adaptive hospital resource management and the optimization thereof. We take a policy-based optimization approach which means that a policy, i.e. a parameterized function, is used to determine an allocation decision given the current state, is optimized in terms of its attributes. The developed policies allow for adaptive resource allocations that improve the optimized allocations obtained in Chapter 4. Moreover, the results show the benefit of incorporating predictions on future resource occupancy to anticipate the effect of allocation decisions taken now on the future. Furthermore, we evaluate the algorithmic settings in order to reduce the computational costs involved in the MO policy optimization approach.

Chapter 6 provides our concluding remarks and discusses possibilities for future work.

A roadmap of this thesis is illustrated by the dependency diagram shown in Figure 1.3. The simulation described in Chapter 2 forms the basis for the computational methods presented in Chapter 3 to Chapter 5. Therefore, the reader is advised to read Chapter 2 as a starter. The prediction method of forward simulation developed in Chapter 3 is applied in Chapter 5, so the reader is advised to consider Chapter 3 for in-depth information on the anticipatory approach in adaptive resource management. The section on supervised learning using neural networks can be skipped at first reading. Finally, Chapter 4 introduces the main concepts of multi-objective optimization and the evolutionary MO optimization approach which is extended in Chapter 5, so the reader is advised to read these chapters in the given order.

1.6 Publications

A paper based on the contents of Chapter 2 appeared as [47]:
A.K. Hutzschenreuter, P.A.N. Bosman, I. Blonk-Altena, J. van Aarle, and J.A. La Poutré. Agent-based Patient Admission Scheduling in Hospitals. In: *Padgham et al., editors, Autonomous Agents and Multiagent Systems – AAMAS 2008*, pages 45–54. IFAAMAS, 2008.

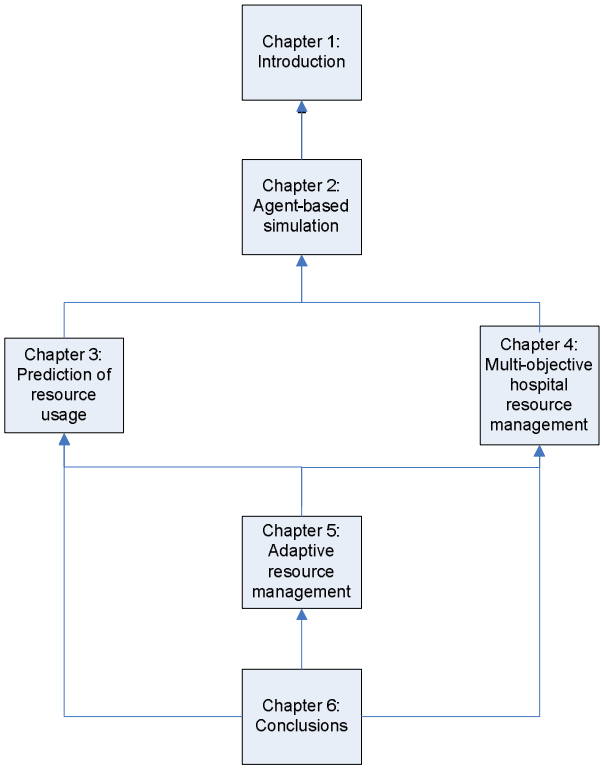


Figure 1.3: Dependency diagram of the chapters in this thesis where the dependencies to Chapter 3 are only with respect to prediction by forward simulation

A paper based partly on Chapter 4 and partly on Chapter 5 was published as [48]:
A.K. Hutzschenreuter, P.A.N. Bosman, J.A. La Poutré. Evolutionary Multiobjective Optimization for Dynamic Hospital Resource Management. In: *Proceedings of the 5th International Conference on Evolutionary Multi-Criterion Optimization – EMO '09*, volume 5467 of *Lecture Notes in Computer Science*, pages 320–334, Springer-Verlag, 2009.

A short version of Chapter 5 with contributions from Chapter 3 will appear as [49]:
A.K. Hutzschenreuter, P.A.N. Bosman, J.A. La Poutré. Enhanced Hospital Resource Management using Anticipatory Policies in Online Dynamic Multi-Objective Optimization. In: *Proceedings of the Genetic and Evolutionary Computation Conference – GECCO 2010*, ACM press, to appear.

Chapter 2

Agent-based simulation for hospital patient flow

In this chapter an agent-based simulation for surgical and emergency patient flows in a hospital is described. The simulation was developed in collaboration with a large university-affiliated hospital in Eindhoven, the Netherlands, and is based on an extensive case analysis, comprising data analysis and interviews. We focus on the coordination of different surgical patient types with probabilistic treatment processes involving multiple hospital units. We also consider the unplanned arrival of other patients requiring (partly) the same hospital resources. The model allows for the assessment of resource network usage as a function of different policies for decision making. Simulation experiments support the validity of our agent-based simulation for decision support. A short version of this chapter has appeared in [47].

2.1 Introduction

As argued in Chapter 1, the hospital domain is an environment of uncertainty and heterogeneity. Patient treatment processes are unpredictable because of stochastic routing between treatment steps, stochastic length of stay (LoS) and possibly stochastic arrival times, e.g. unexpected emergency patients, cf. Chapter 1, Section 1.1.3. Moreover, patient treatment processes often involve several hospital units. Often, resources, e.g. at the ICU, are shared by multiple treatment processes. The actual flow of inpatients through the network of hospital units, however, depends on the available resources at the different units. Thus, actual patient pathways may deviate from the medically indicated treatment processes, patient transfers may be

deferred, and patient admissions and operations may need to be canceled. Due to the stochastic patient processes and the actual patient flow being the result of resource availability, an analytical evaluation of such a problem setting is not feasible. Furthermore, changing the structure of the patient pathways or the underlying probability distributions is non-trivial in an analytical model. In order to gain insight into the functioning of such a complex and dynamic system we developed a simulation for this problem setting. In addition to providing insight, the simulation allows for predicting effects of changed system or environment variables (so called what-if scenarios) that would otherwise be unpredictable, e.g. changing the parameters of the units' transfer and admission policies or hospital resource allocations.

Hospitals often show a distributed organizational structure [26, 59, 81]. They are divided into several autonomous hospital units that are each associated with a medical specialty. Schedules of shared resources, like operating rooms, are managed locally by the units each applying their own (medical) priorities and preferences. Thus, patient scheduling in hospitals has strong decentralized features, cf. Section 1.1.1. In order to obtain a simulation that realistically represents the problem domain and allows for accurate what-if evaluations, the simulation model should not only incorporate the stochastic features related to patient care of different patient groups, but also reflect the distributed decision making by the different hospital units. Therefore, an agent-based simulation is a natural candidate to model hospital reality.

The agent-based simulation was developed in a cooperation between academia and the Catharina Hospital Eindhoven (CHE), the Netherlands. As outlined in Section 1.3, we base the simulation model on an extensive case analysis at the cardiothoracic surgery (CTS) department at the CHE. The CHE is a large university-affiliated hospital in Eindhoven with state-of-the-art intensive care and CTS medicine. The case study comprises an extensive data analysis and several interviews with experts from the CTS unit and the ICU. Specifically, we consider CTS-patient flows and their interaction with other surgical and emergency patient flows. The different hospital care units involved in the treatment of these patients are represented each by an autonomous agent in the simulation. The following features are included in the simulation:

- different patient characteristics that influence the patients' priority and pathway in the hospital,
- patient treatment processes with stochastic routing and LoS,
- multiple hospital care units with individual scheduling policies, and

- limited and uncertain resource availability due to the inflow of other surgical patients and the arrival of emergency patients.

To the best of our knowledge, this is the first agent-based simulation for hospital patient scheduling that includes the features above and that is based on real hospital data and scheduling practice.

The remainder of this chapter is organized as follows. First, we discuss related work in Section 2.2. The design of the agent-based simulation is described in Section 2.3 with the architectural structure, the decision model of the agents, the patient pathway model and the case study inputs. Then, the simulation experiments that were performed for evaluating and validating the simulation are reported in Section 2.4. Finally, in Section 2.5 we provide our conclusions.

2.2 Related work

As mentioned in Chapter 1, Section 1.4, relevant related work on patient scheduling and simulation in hospitals can be found in the Operations Research and Management literature and AI literature on agent-based modeling and optimization. Typical planning problems considered in the Operations Research literature relate to single care units. Especially the problems of operating room scheduling (e.g. [18, 23, 39]), the allocation of hospital beds to a care unit (e.g. [35, 40, 55, 58, 80, 89, 97]) and the scheduling of diagnostic facilities (e.g. [52, 57, 73, 74]) have attracted major research interest. In our work, we focus on complex treatment processes that involve multiple hospital units, i.e. the OR, ICU and nursing wards. The work reported in [60] and [1, 99] provide theoretical results for bed utilization levels for patient treatment processes with deterministic routing. We offer a more operational approach which can deal with stochastic treatment durations and routing. Moreover, our agent-based simulation approach is very flexible and can be easily adapted to other settings matching structural hospital features. The simulation model presented in Harper and Shahani [40] facilitates the evaluation of aggregated bed allocation policies. Our approach allows for an in-depth analysis of allocation strategies also on the level of different hospital units. Additionally, the effect of (small) changes in bed allocations can be evaluated using the agent-based simulation tool. The simulation study in VanBerkel [93] focuses on waiting time reduction and capacity planning for partly deterministic patient flows of the general surgery specialty. We consider complex patient pathways involving multiple postoperative care units, that are also used by other specialties' patient

pathways, and stochastic routing. In Groot [37] different surgical admission policies are evaluated in a setting that involves the OR and a general pool of postoperative care beds with a stochastic demand for care and deterministic treatment processes. In our model we consider stochastic patient processes and a hospital setting with multiple postoperative care units that are required by several patient pathways. The work in Blake and Carter [9] addresses the optimization of the overall patient case mix on a strategic level, whereas our approach is on an operational level of control for patient (admission) scheduling. The stationary Markov chain model described in Kolesar [58] is used to optimize admissions given the estimation on available resources while our approach explicitly includes fluctuations in hospital resource utilization due to the stochastic patient treatment processes.

Fewer approaches on agent-based patient planning and scheduling have been proposed in the AI literature. In [26] the issue of conflict handling in patient task scheduling is studied. However, the dynamics of the problem, like stochastic treatment durations and stochastic routing, as well as different patient characteristics are not considered. Random treatment durations and routing between treatment steps are, however, very important to consider because they perturb the hospital units' schedules. The approach described in [75, 76, 77] for patient task planning is based on medical wellness functions of patients. In our work, we did not include medical wellness functions for patients since the utility elicitation of representative wellness functions is considered time-consuming, data collection is resource-intense and is still an open problem in epidemiological research [46]. Moreover, the solution does not scale sufficiently and does not consider resource constraints for inpatient planning and the stochastic patient care features incorporated in our approach. Multiple appointments in an outpatient setting have been studied in [94]. Their approach assumes a predefined treatment path which does not hold in our problem setting. Also, no stochastic appointment lengths were considered.

2.3 Simulation model

In the following section the design of the agent-based simulation is presented. Our system of concepts follows the terminology of Fishwick [34]. The context of the simulation is a network of hospital care units through which patients of different types are flowing according to their pathways, cf. Chapter 1, Section 1.1.3. As described in Section 1.3, we designed and implemented a simulation for this setting based on a case study at the Catharina hospital

Eindhoven (CHE) using the paradigm of software agents to represent the different care units involved in the patient treatment.

For the analysis and design of the agent-based simulation, the methodologies in [29] and [70] were taken into account. During the development phase, the model and functionality were frequently discussed with hospital planners and managers at the CHE. The resulting model was approved by the CHE domain experts.

In the following, we first discuss the requirements and goal of the simulation. Then, we present an overview of simulation with the different agents involved, their decision models and a model of the patient flows.

2.3.1 Requirements & goals

The goal of the simulation is to realistically model and simulate patient flows in a hospital setting. On that account, the model design should capture the autonomy of hospital care units. Specifically, the simulation model should facilitate that the different care units autonomously initiate and evaluate patient admissions, schedule medical procedures and arrange patient transfers & discharges according to the patients' pathways. Moreover, the model should allow for care units to react flexibly to bottlenecks arising from high resource occupancy in their units through dynamic adjustment of patient transfers under some medical restrictions. Furthermore, the simulation should be flexible and adjustable to other patient pathway and hospital settings. Considering the execution of the simulation model, the model should be executable on a single PC and feature a reasonable runtime.

2.3.2 Architecture of the simulation model

Figure 2.1 provides an overview of the agent-based simulation. The simulation is composed of three major sections: the case inputs, the simulation and its output. The figure also includes the relation of the simulation to three adaptive computational models for predicting future resource usage and optimizing dynamic resource management and admission control which will be discussed in the remaining chapters of this thesis. Below we discuss the main parts of our simulation in more detail.

Case study inputs The inputs of the simulation can be divided into information concerning patient care and general conditions for hospital operations. The patient information comprises the relevant patient groups and their pathways, patient priorities and waiting lists. The model that is used

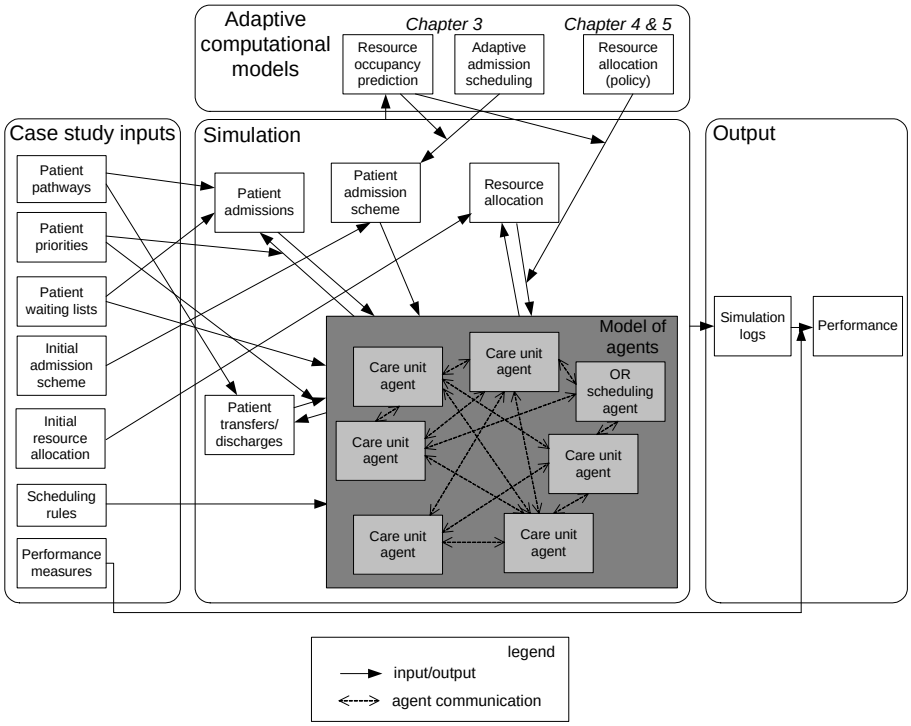


Figure 2.1: Overview of the agent-based simulation

in the simulation for the patient pathways will be discussed in detail in Section 2.3.4. The general conditions for operations in the simulation include an initial allocation of resources to the different units, a basic admission scheme, scheduling rules and output measures for the performance of the simulation. The case inputs of the simulation are obtained from an extensive data analysis of historical data and interviews with domain expert from the CHE. The case study and the derived inputs are reported in further detail in Section 2.3.5.

Resource allocation In accordance with Section 1.1.3 the resource allocation in our simulation specifies the number of resources, i.e. half-day OR time slots and hospital beds, that are allocated to the different care units and specialties. Moreover, the resource allocation indicates the temporal availability of the allocated resources. The resource allocation may be adjusted online based on a resource allocation policy and information obtained from the care unit agents in the simulation. The online adjustment of resources will be further discussed in Chapter 4 on the optimization of hospital

resource allocations.

Patient admission scheme As introduced in Section 1.1.3 the admission scheme controls the flow of patients into the system of hospital units represented in the simulation. The scheme determines the (maximal) daily number of patients to be admitted to the hospital which also depends the availability of elective surgical patient groups on the waiting lists and their medical priorities, e.g. [75].

Patient admissions, transfers and discharges The simulation of the patient pathways generates events to indicate the arrival or possible transfer of a patient. Patient arrivals initiate the internal decision making of the corresponding agent to evaluate the admission. Patient transfers trigger the interaction between the involved agents. As explained in Section 1.1.3, the actual patient admissions, transfers and discharges are the result of the agent communication.

Model of agents The agent model comprises two types of agent: OR scheduling agents and care unit agents. An OR scheduling agent represents a surgical specialty and is responsible for managing the schedule for the allocated OR time slots. Care unit agents, in abbreviation also referred to as unit agents, act on behalf of postoperative and critical care hospital units. The unit agents coordinate patient transfers with other agents based on the patients' pathways and the available resource capacity. Furthermore, unit agents decide upon patient admissions on the basis of the admission scheme and the available resources. The agent mapping and the agents' roles were identified by use cases obtained from the CHE case analysis.

The different agents negotiate patient transfers in the simulation through message exchange indicated by dashed arrows in Figure 2.1.

The internal decision making model of the OR scheduling and care unit agents are described in detail in Section 2.3.3, respectively.

Output The system offers logging possibilities for actual patient admission, transfer and discharge decisions which is used to determine different outcome measures. The outcome measures include

- the patient throughput, i.e. the number of patients that completed their treatment process, for the different patient groups,
- the number of patients per group treated at the different units,

- the number of rejected patient admissions at the different hospital units,
- the number of rejected patient transfers per patient group at the different hospital units,
- the number of canceled surgery sessions for the different patient groups due to unavailable postoperative care beds, cf. Section 1.1.3,
- the frequency and duration of back-up capacity usage at the different units for the different patient groups, and
- the total costs for regular resource capacity at the different hospital units.

Moreover, the simulation model offers several statistics for evaluating the different output measures. In accordance with the hospital management at the CHE, the performance measures of interest are determined by the mean and standard deviation of the patient throughput, the costs associated with total regular bed capacity and unused OR capacity as well as the accumulated back-up capacity usage.

Decision variables In accordance with the hospital managers from the CHE, the following variables are considered as free decision variables in the simulation:

- the number of allocated beds and ORs,
- the patient admission scheme, and
- the parameters of the unit agents' patient (re-)transfer policies.

2.3.3 Decision model of agents

Patient transfers are negotiated among the different agents in the simulation through message exchange. In their decision making process whether to accept or reject requested patient transfers the agents apply their individual scheduling policies and decision criteria. In the following we first give a general description of the agents' communication and decision making process. Then, we address the specific decision making models of the different types of agents.

General description

In the communication protocol employed by the agents, a patient transfer is proposed from the agent that represents the unit where the patient is currently admitted to the indicated care unit's agent. The applied agent evaluates the proposed transfer based on its admission policies and other decision criteria. The agent then returns a message to inform the applying agent how many of the proposed patient transfers are accepted, if any. If a patient transfer is accepted, an additional message is sent to inform the accepting agent of the actual patient transfer. Thereafter, the patient is admitted to the care unit and assigned to a resource.

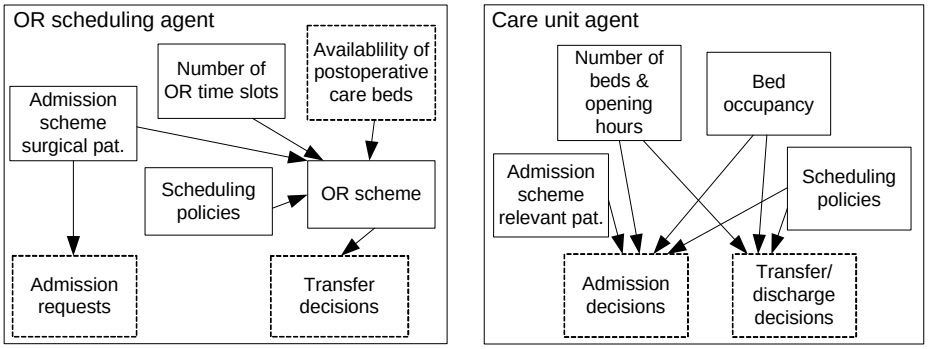
In general, patients are admitted to a hospital unit only if resources are available, excluding available back-up capacity. In our simulation some exceptions to this rule are made for patients that are already admitted to the hospital which is inspired by the scheduling policies employed at the case study hospital which is described in more detail in Section 2.3.5. In line with current hospital practice this means that if more patients are proposed for transfer than resources are available, a choice must be made which patients are admitted. This decision is typically based on a multitude of clinical variables. As explained in Section 1.3 the medical choice for assigning surgical and emergency patients to available resources is represented by a stochastic process that first randomly selects patients with the highest priority level, i.e. emergency status. Any remaining available capacity is then assigned randomly to the other patients proposed for transfer. Due to the generic setup of the simulation also a more elaborate choice model can potentially be incorporated in a straightforward manner.

In the simulation the agents continually check whether a patient is fit for transfer or discharge. The eligibility for transfer and discharge is based on the LoS, i.e. if the LoS of the patient at the specific unit has elapsed[1]. If a patient is admitted to a hospital unit that is not indicated by the patient's pathway due to resource shortage at the concerned unit, cf. Section 1.1.3, the corresponding care unit agent may also propose later on to re-transfer the patient to the care unit originally indicated. The LoS for the indicated care unit is then reduced by the time the patient spent at the transferring unit. Also, the agents attempt to accommodate patients that have been admitted to back-up capacity to a "regular" bed if one becomes available through another patient's transfer or discharge.

Figure 2.2 provides a graphical representation of the decision models of

[1]In reality the eligibility for transfer is a medical input that is assessed by the medical staff.

the different types of agent in the simulation which will be described in more detail below.



(a) OR scheduling agent (b) Care unit agent

Figure 2.2: Decision models of agents in agent-based simulation for the (a) OR scheduling agent and (b) care unit agents (dashed lines indicate agent communication)

OR scheduling agent

The task of the OR scheduling agent is to schedule surgical patients to OR time slots. This comprises the following actions and decisions:

- prior to surgery: initiate the admission of the surgical patients to fill available OR time slots of the respective specialty,
- on the day of surgery: determine OR scheme & requests for pre- and postoperative transfers to the corresponding care unit agents.

The number of patient admissions that are initiated is derived from the admission scheme of the relevant patient types. Due to the cost pressure, surgical patients are commonly admitted to the hospital only few days prior to their planned surgery. In our simulation a fixed preoperative LoS of one day is assumed.

The agent's scheduling policies determine the OR scheme which is initially derived from the admission scheme of the day prior to surgery. The scheduling policies incorporate medical and organizational rules imposed by the hospital context and specify. For example, the rules comprise constraints on allocating OR time slots to type of patients, e.g. time slots in the morning are reserved for children or patients with postoperative indication for transfer to the PACU which is opened only for a limited period

of time. Moreover, the scheduling policies require that a postoperative care bed is available and reserved at the indicated care unit in order to perform a surgery. The meaning is as follows. If a postoperative care bed is available, a patient with corresponding patient type is always scheduled for an available OR time slot. However, if postoperative care beds are unavailable, the OR scheduling agent reduces the number of corresponding surgeries in OR scheme accordingly.

Based on the OR scheme, the OR scheduling agent requests patients planned for surgery to be transferred from their current units to the OR theater. Also, the OR scheduling agent requests the postoperative patient transfers to the indicated care units.

Care unit agents

Care unit agents perform the task of coordinating patient transfers with the different agents involved in the different patient pathways. Regarding patient admission, transfer or discharge decisions the following factors are taken into consideration: the current bed occupancy, the number of allocated resources, the agent's scheduling policies, the admission scheme and the messages exchanged with the other agents.

The agents' scheduling policies affect the admission and transfer decisions as follows:

- whether to reserve resources(s) for specific patient groups,
- whether to accept patient admissions & transfers and to which extent depending on the available resources (excluding reserved capacity),
- whether and how back-up capacity is allocated to patients,
- which patients are selected to undergo surgery if requested by OR scheduling agent,
- whether an alternative patient transfer is arranged if the originally indicated care unit agent rejects the transfer, and
- whether patient re-transfer is attempted.

As explained above, a care agent only accepts a patient transfer or admission if a resource is available to assign the patient to. Following the clinical rule at the CHE, patients in more severe clinical condition are given priority to be selected for transfer or admission. Also, some care unit agents may use back-up capacity in order to accommodate patients of specific patient types.

Concerning preoperative patient admissions requested by OR scheduling agent(s), a care unit agent may account for eligible preoperative patients that are already admitted due to previously canceled surgeries. Based on organizational rules at the CHE, in our model patients remain admitted to the corresponding care unit if their surgery has been postponed.

For selecting surgical patients for available OR time slots the care unit agents apply a random choice. This is a representation of reality where this decision is taken by the specialty's medical staff. Through the random choice it may happen in the simulation that a patient remains at the care unit for a "long" time. In reality the chances for being selected for an OR slot increase with increasing time spent at a care unit. However, since we distinguish the patient types this does not become a limiting factor for the remaining treatment process.

If a patient transfer is rejected a care unit agent may arrange for an alternative patient transfer. This means that a (temporary) patient transfer is requested to another unit that is not indicated by the corresponding patient pathway. According to medical rules, the first option is restricted to care units of equal or higher care level than the unit originally indicated. The patient pathway is then adjusted accordingly. If the transfer to another unit of equal or higher care level is not possible, the patient has to remain admitted to the corresponding unit. Depending on the scheduling policy, this procedure is repeated until the patient transfer is accepted or the LoS has elapsed. It should be noted that if patients remain admitted at their current unit, this may especially affect surgical patient flows to the OR since appropriate postoperative care resources must be available and reserved for patients before undergoing surgery.

If a care unit agent has accepted the (temporary) transfer of a patient who has a different, possibly lower, care indication, the agent may decide to initiate the re-transfer of the respective patient to the originally indicated care level. This decision may depend on the current resource utilization, cf. Section 2.3.5.

The agent-based model reflects the complex features of the hospital domain in a detailed and realistic way. The case study and the relevant inputs are described in detail in Section 2.3.5. The experimental evaluation and validation of the simulation is presented in Section 2.4.

2.3.4 Model of patient pathways

This section describes the model of patient pathways that is used in the simulation. The model is a formalization of the model described in Chap-

ter 1. As discussed in Section 1.1.2 and Section 1.1.3, we consider patient groups with similar resource consumption during their treatment. The set of patient groups that can be distinguished on the basis of the resource need is denoted by Θ. Let U denote the relevant hospital units involved in the patient treatment processes to be considered. All possible pathways of a patient type $g \in \Theta$ are modeled as a probabilistic graph [10]. The graph is given by the tuple $G^g = (U^g, A^g, P^g)$, where $U^g \subset U$ denotes the hospital care units involved in the treatment process and A^g represents the set of arcs between units $u \in U^g$, i.e. the possible adjacent treatment operations. The length of stay (LoS) of a patient of group $g \in \Theta$ at hospital unit $u \in U^g$ is modeled as a random variable, LoS_u^g, that follows a discrete probability distribution $P^{LoS_u^g}$. P^g is the set of conditional probability distributions defined on A^g with

$$P^g = \{Pr(v|u, g, t)|u, v \in U^g, (u, v) \in A^g, t \geq 0\} \text{ for } g \in \Theta. \tag{2.1}$$

$Pr(v|u, g, t)$ represents the probability that care provided by unit v is required given that a patient of type g has been admitted to unit u for t time units. The possible patients'discharge destinations from the hospital are indicated by $o \in U$ which comprise home or other care facilities, but also mortality.

Consider for example the patient pathway illustrated in Figure 2.3 where a surgical patient treatment graph of a patient type g is depicted. The care units involved in the treatment process are $U^g = \{\text{OR}, \text{HC}, \text{Ward}\}$, where HC denotes the High Care, cf. Table 1.1 on page 4. The units are represented as nodes in the patient graph with associated LoS indicated by the node, $P_u^{LoS^g}$, $u \in U^g$. The arcs of the graph represent the clinical decision between adjoining treatment steps. Here, all patients of type g undergo surgery at the OR, then a patient may either require HC care after surgery or ward care which is determined by the conditional probabilities $Pr(HC|OR, g, t)$ and $Pr(Ward|OR, g, t) = 1 - Pr(HC|OR, g, t)$. All patients admitted to the HC subsequently require ward care and finally all patients are discharged from the ward after completed treatment.

As outlined above and in Section 1.1.3, the actual flow of patients between hospital units is the result of negotiation between the corresponding units for which the decision models of the different agents discussed above are taken into account. The patient pathways thus specify the patient flows in a situation of unlimited resource availability. Possible destinations of patients' discharges from the hospital are home or other care facilities, but also mortality.

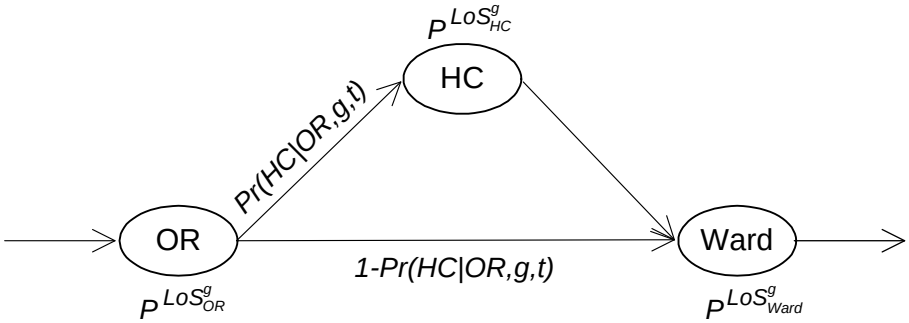


Figure 2.3: Illustration of probabilistic graph model of a patient pathway

2.3.5 Case study

The simulation instance that is used throughout the remainder of this thesis is based on an extensive case analysis at the CTS department at the CHE. Below, we describe the different case inputs for the simulation.

Patient pathways and priorities

The following patient pathway description is based on numerous expert interviews in combination with an extensive data analysis. In the CTS case study, the relevant care units are included in the set U that is given by U={CTS-OR, IC, IC-HC, MC, CCU, CTS-HC, CTS-PACU, CTS-ward, o}, cf. Chapter 1, Section 1.1. Here, the prefix CTS indicates that a hospital unit is (partly) dedicated to CTS patients, e.g. OR time slots assigned to the CTS specialty. The High Care (HC) unit is subdivided into IC-HC, which is shared by different surgical specialties, and CTS-HC which occasionally allows other patients as well. o summarizes the possible destinations of a patient's discharge from the hospital, cf. Section 2.3.4.

In the CHE case study for the CTS, four types of patient pathways (type I to IV) were identified. Type I and II patients are CTS patients, for whom the first postoperative care is indicated as CTS-HC and CTS-PACU, respectively. The decision for a type I or II pathway is based on a preoperative assessment of the patient's clinical condition. The type III pathway corresponds to the treatment process of emergency patients who arrive unexpectedly. The type IV patient path represents the inflow of other surgical patients in the system. Other surgical patients comprise patients from amongst others general surgery, orthopedics and gynaecology. These patient flows are aggregated because of the strategic focus on cardiothoracic surgery at the CHE and the resulting relatively small number of patients for

other surgical specialties.

The pathway of type I patients is depicted in Figure 2.4. Here, type I patients undergo surgery during the OR time slots allocated to the CTS specialty, denoted as CTS-OR. After surgery, they are admitted to the CTS-HC for recovery from surgery and are expected to return to the CTS-ward on the following day. The schedule of the ward round at the CTS-HC determines the fixed point in time when patient transfer decisions are taken. This implies that the LoS at the CTS-HC can be considered as deterministic and t is irrelevant in (2.1). Complications require an admission to IC or MC for 15% of the type I patients. Patients admitted to IC or MC are subsequently transferred to the CTS-ward. If type I patients no longer require medical care and monitoring in the hospital, they are discharged and leave the system. Complications requiring re-admission or re-operation can be easily incorporated in our model. In the considered CTS case study, however, they were irrelevant because they occur only exceptionally (in about 0.6% of the cases). Figure 2.5 shows the four types of patient pathways

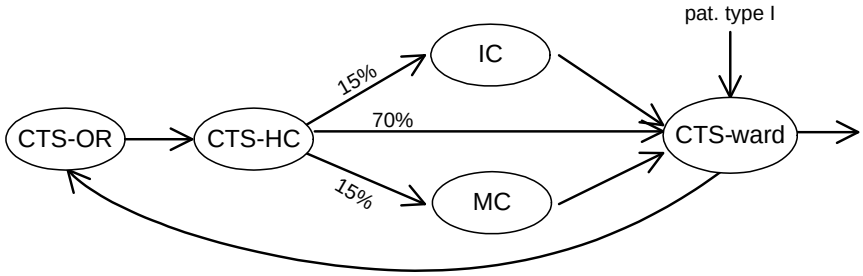


Figure 2.4: Representation of type I patient pathway[1]

and their interference. By dashed arcs, the possible pathways of type II patients are depicted. Type II patients follow a fast-track variant of the type I path. Postoperative recovery and care is performed at the CTS-PACU and type II patients are expected to return to the CTS-ward on the same day of surgery. Similarly to type I patients, the LoS at the CTS-PACU can be considered as deterministic as type II patients have to be transferred to another care unit at the closing of the CTS-PACU. Severe complications that require admission to IC or MC occur but rarely and the corresponding routing probabilities to IC and MC are given as 5% and 15%, respectively[1] . The postoperative and critical care units involved in the CTS patient treatment processes are also partly required by other patients. The corresponding resource requirements are depicted by the type III and IV pathways in Figure 2.5. In our research we focus on the possible interference between the

different patient pathways. For this reason, we restrict the model of the type III and IV pathways to the IC, IC-HC, CTS-HC and MC unit as the preceding and successive treatment steps of type III and IV patients involve other dedicated resources that are not required by CTS patients. Therefore, type III patients arrive at the IC or MC while type IV treatment primarily involves the IC-HC. If IC-HC beds are scarce, type IV patients may alternatively be routed for admission at the IC or CTS-HC[1]. In this model we

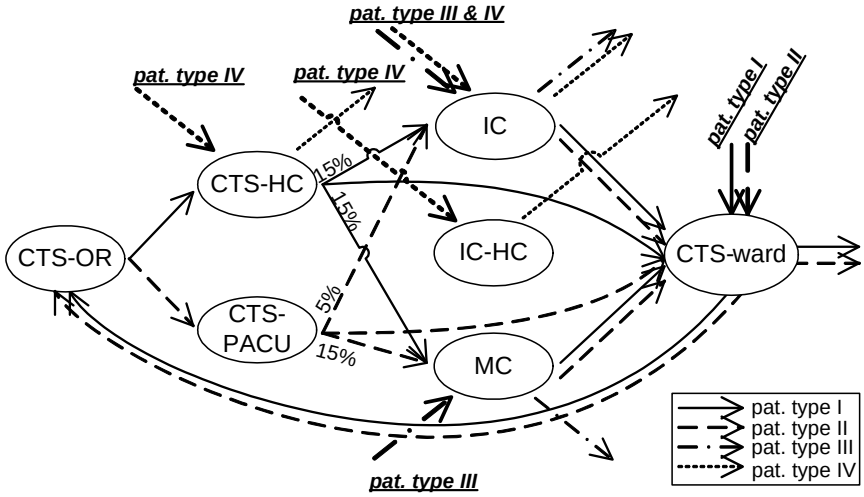


Figure 2.5: Interference of CTS, other surgical and emergency patient pathways[1]

distinguish two levels of patient priority: the "emergency" priority status which is assigned to type III patients and type I and II patients with IC indication after surgery and the remaining "regular" patients. In case of resource scarcity the CHE units primarily accept "emergency" patients while "regular" patients are eligible for alternative routing.

Resource allocation

For the treatment of the CTS patient groups at the CHE four half-day OR sessions are allocated to the CTS specialty. For the CTS, a half-day OR session corresponds to one surgery that can be performed. The early OR slots are assigned to type II patients which is required by the design of the CTS patient pathways. As described in Section 2.3.5, the CTS-PACU beds are only opened for a limited time and the type II patients require a postsurgical recovery time before returning to the CTS-ward. The number of postoperative care beds at CTS-PACU and CTS-HC equals the number

of OR slots allocated to the CTS specialty. Beds at the CTS-PACU and CTS-HC are opened for a limited time window. Moreover, there are four IC-HC and MC beds and 11 IC beds for acute care and monitoring. For the revalidation at the CTS-ward 35 beds are allocated. The hospital beds at IC, IC-HC, MC and the CTS-ward are opened 24 hours, 7 days per week which is indicated as 24/7. The resource allocation and availability of resources at the CHE is also summarized in Table 2.1.

Unit	Number of resources	Resource availability
CTS-OR	4 ORs	Mo-Fr 8h00-12h00
CTS-OR	4 ORs	Mo-Fr 12h00-17h00
CTS-PACU	4 beds	Mo-Fr 12h00-22h00
CTS-HC	4 beds	Mo 10h00-Sa 10h00
IC	11 beds	24/7
IC-HC	4 beds	24/7
MC	4 beds	24/7
CTS-ward	35 beds	24/7

Table 2.1: Number and availability of allocated resources at the different care units in CHE case study

Agents' policies for agent-based simulation

In the simulation, the OR scheduling agent represents the CTS specialty. The care unit agents act on behalf of the IC, IC-HC, MC, CTS-HC and the CTS-PACU unit and the CTS-ward. The implemented scheduling policies of the different agents are summarized in Table 2.2 on page 46. Below the policies are described in further detail.

Communication moments The CHE patient pathways and resource availability constraints result in a number of fixed time points for admission and transfer communication among the OR scheduling and several care unit agents in the simulation. A time line with the timely fixed decision and communication moments for the involved agents is depicted in Figure 2.6. Except for the fixed decision and communication moments, care unit agents initiate patient transfer communication if a patient is eligible for transfer, cf. Section 2.3.3.

OR scheduling agent The OR scheme of the OR scheduling agent for the CTS specifies the number of type I and II patients to be scheduled to

Agent	Scheduling policy
OR Scheduling agent	request transfer of type II and I from CTS-ward to OR as specified in OR scheme at time $\mathbf{S}_{\mathsf{CTS\ II}}$ and $\mathbf{S}_{\mathsf{CTS\ I}}$, respectively (cf. Figure 2.6); reserve beds at CTS-PACU and CTS-HC prior to surgery; if informed that insufficient beds are available, cancel surgeries accordingly; based on OR scheme for following day, inform CTS-ward on required number of type I and II patients at time $\mathbf{A}_{\mathsf{I+II}}$; send transfer requests to CTS-PACU and CTS-HC agents after completed surgery of type II and I patients, respectively
CTS-PACU agent	send transfer requests to hospital unit indicated for admitted patients at time $\mathbf{T}_{\mathsf{CTS\text{-}PACU}}$
CTS-HC agent	send transfer requests to care unit agents at time $\mathbf{T}_{\mathsf{CTS\text{-}HC}}$; if transfer is rejected by all possible care unit agents, patients remains (temporarily) at CTS-HC (and is repeatedly proposed for transfer, depending on the settings of the re-transfer policy); on request of the OR scheduling agent reserve beds for postoperative type I patients; accept admission of type IV patients if beds are available; possibly retry patient transfer for type I to MC and re-transfer type IV patients to IC-HC
IC agent	admit all type I & II patients with IC indication, possibly use back-up capacity if IC beds are scarce; other patient admissions are accepted by random choice over patients contained in transfer proposal, retain one bed for type III patients; retry transfer of CTS patients to MC and type IV patients to IC-HC, respectively, depending on the settings of the re-transfer policy
IC-HC agent	if insufficient beds are available for requested type IV admissions, send admission request to IC (and if necessary the CTS-HC) agent; if unsuccessful, reject admission; admit other patients proposed for transfer by random choice to free beds; retry transfer of CTS patients to MC, depending on the settings of the re-transfer policy
MC agent	admit all patients from CTS-PACU; if MC beds are scarce, use back-up capacity; admit other patients proposed for transfer by random choice to free beds
CTS-ward agent	admit all postoperative patients; the number of preoperative admissions depends on the following day's OR scheme accounting for previously admitted patients whose surgeries have been canceled; if OR scheme is adjusted due to unavailability of postoperative care beds, randomly select patients of corresponding patient type; if ward beds are scarce, use back-up capacity

Table 2.2: Scheduling policies implemented in agent-based simulation system

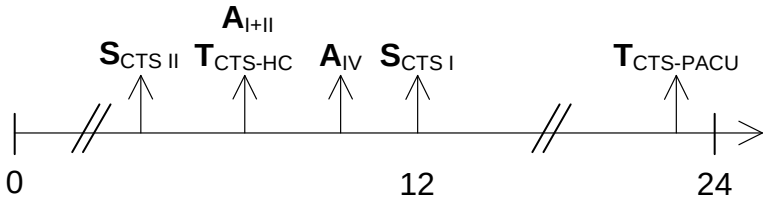


Figure 2.6: Time line for the fixed decision moments and communication: S - schedule surgery, T - transfer, A - admission and respective agent/patient group

the allocated OR time slots, i.e. 2 half-day sessions for each of the 4 ORs. As described above, the early OR slots are assigned to type II patients due to the design of the CTS patient pathways.

IC agent At the CHE, type I and II patients with IC indication are considered like emergency patients and are assigned the same priority. In accordance with CHE practice, they are always admitted to the IC if required. If free IC beds are scarce, the IC agent may use back-up capacity to accommodate the emergency patients which is accounted for in the system's performance. At the same time, beds may be reserved for arriving type III patients. At the CHE, this policy is applied for one IC bed. If the admission of type III patients from outside the hospital system is rejected by the IC agent this is accounted for in the performance measures. In this case patients are admitted to another hospital which, however, is beyond the scope of our work and consequently left out of our model.

MC agent We analyzed at the CHE that the MC agent always accepts transfer proposals from the CTS-PACU because CTS-PACU beds are closed at 22pm. If MC beds are scarce, patients will be accommodated to back-up capacity.

CTS-HC agent The alternative patient routing possibilities are illustrated in Figure 2.7 for type I patients. Here, the possible consecutive patient path is depicted by bold arrows. If a type I patient, that is currently admitted to the CTS-HC with a MC indication, is rejected for transfer to the MC, the CTS-HC agent approaches the IC-HC agent which normally is not intended for in the type I patient pathway, cf. Section 2.3.5 and Figure 2.4. If the proposed transfer is accepted by the IC-HC, the patient is transferred to the higher care level. Otherwise, a care agent of the next higher care level is approached, here the IC agent. If the transfer to the IC unit is not

possible, the second rule applies and the patient has to remain admitted to the CTS-HC until closing time or indicated transfer to the CTS-ward. Possibly, the patient is repeatedly proposed for transfer to the MC, this depends on the settings of the agent's re-transfer policy. It should be noted that this second option may affect the type I patient flow to the OR since appropriate postoperative care resources must be available and reserved for patients before undergoing surgery.

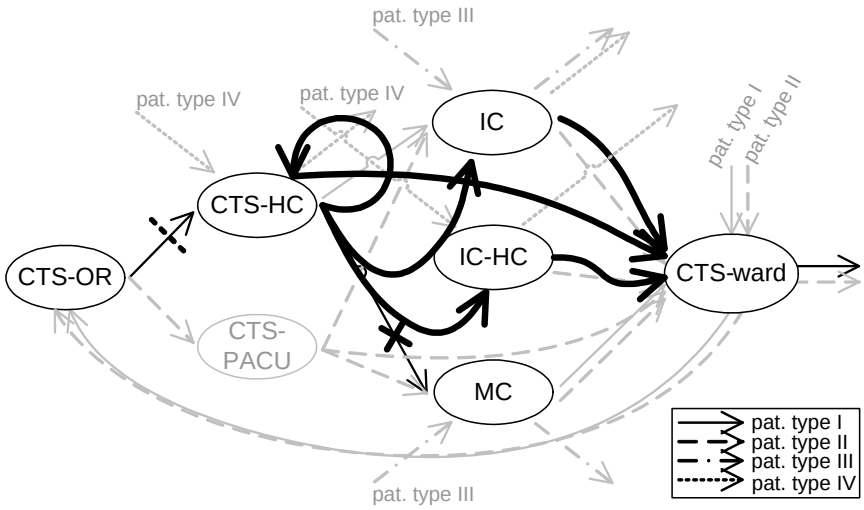


Figure 2.7: Current practice for alternative patient routing of type I patients

IC-HC agent Similarly to the CTS-HC agent, alternative patient routing may also be applied by the IC-HC agent for the admission of type IV patients. At the CHE, first the IC agent is approached, followed by the CTS-HC agent. If alternative routing is not feasible, type IV patient admissions are rejected. A rejected admission may affect the corresponding surgical specialty's OR schedule and may cause blocking at the dedicated nursing ward. Since our focus is on the interference between type I, II, III and IV patient flows, only the relevant care units in the type IV patient pathway are considered. The effects of rejected type IV transfers are not accounted for in the system's performance.

CTS-ward agent Given the admission requests for type I and II patients, the CTS-ward agent schedules the patients' admissions to the CTS-ward for the following day while accounting for previously canceled surgeries. In

our model patients remain admitted to the CTS-ward unit if their surgery has been postponed. The CTS-ward agent randomly selects type I and II patients for the available OR time slots. Postoperative type I and II patient transfers are always accepted by the CTS-ward agent, if necessary back-up capacity is used to accommodate the patients.

Data analysis

In addition to several expert interviews at the ICU and the CTS department of the CHE, we conducted an extensive data analysis to obtain real life patient pathway data for the simulation. The relevant anonymized patient admission data was obtained by combining pathway data from multiple, partly unit-specific, hospital information systems at the CHE.

Data inclusion and initial analysis The inclusion criteria were set to select all patient admission records of CTS, emergency and other surgical patients with postoperative ICU care that completed their treatment at the CHE[2] in the year 2005. The obtained data was inspected on an aggregated level and erroneous data were corrected, if possible, or otherwise excluded. For example, if the LoS at the CTS-PACU unit exceeded the official working hours due to delayed patient transfers or entry into the system. Also, outliers with a LoS of more than 50 days were removed as it appeared in discussions that very long LoS often indicated dateless patient transfers.

Results The relevant input parameters, i.e. the parameters determining the LoS distribution and the conditional routing probabilities, of the different patient pathways introduced in Section 2.3.5 are given in Table 2.3.

Patient routing The proportion of type I patients in the CTS patient population amounts to 60% which corresponds to preliminary estimates obtained from CHE domain experts. The type I patient routing probabilities after the patients' stay at the CTS-HC derived from the data amount to 0.1 and 0.2 for the MC and IC (including IC-HC), respectively. Due to the current practice of alternative patient routing in case of resource scarcity at the CHE, cf. Figure 2.7, the routing probabilities were set to 0.15 for MC and IC, respectively, corresponding to the medical transfer indications on

[2]This criterion excludes deceased CTS patients who amount only to a small proportion of the total patient volume, cf. Section 2.3.5.

Patient group	Unit	LoS (hours) mean±stdev	Routing prob.
Type I	CTS-HC	15 ± 0	-
	IC	48.48 ± 54	0.15
	MC	24.48 ± 38.52	0.15
	CTS-ward	120 ± 22.08	0.7
Type II	CTS-PACU	6 ± 0	-
	IC	42 ± 57.12	0.05
	MC	10.32 ± 22.08	0.15
	CTS-ward	120 ± 22.08	0.8
Type III	IC	89.48 ± 200.82	-
Type IV	IC-HC	34.94 ± 68.51	-

Table 2.3: Parameters of patient pathways

the basis of CHE expert knowledge. The type II routing probabilities are not affected by the transfer practice and were therefore based on the data analysis and amount to 5% and 15% for postoperative MC and IC transfers, respectively. No correlation could be found between the LoS and subsequent patient routing, therefore the conditional routing probabilities in (2.1) will be considered independent of t.

Patient LoS For modeling the patient LoS data we considered typical LoS models proposed in the literature were considered, e.g. in [41, 63, 65]. Typical LoS distributions are skewed, i.e. asymmetric, and contain outliers. Skewness of LoS distributions is characterized by a long-sided tail on the right, which means that the probability mass is concentrated on small values of the distribution with relatively few high values as outliers. Three most widely used models for sampling patients' LoS are Lognormal, Gamma and Weibull distributions [63]. Additionally, we evaluated Poisson, Beta, Erlang, Normal, Exponential, Triangular and Uniform distributions and their fit using the Kolmogorov-Smirnov goodness-of-fit test [45]. Overall, Lognormal and Gamma distributions appeared to best fit the CHE LoS data. We chose Lognormal distributions as their use is simple and fast. Moreover, Gamma distributions did not result in significantly different simulation results in the basic setting. In accordance with expert opinion, alternative patient routing according to the rules explained in Section 2.3.5 does not affect the LoS of a patient. Also, no correlation could be found between the LoS at the different units. Therefore, the Lognormal distribution parameters were estimated independently using the method of moments [45].

Since the focus of the simulation are the surgical and emergency patient flows rather than the surgery scheduling, we assume a constant surgery duration, i.e. a half-day session corresponds to one surgery. Based on CHE practice, a constant preoperative LoS at the CTS-ward of one day is assumed.

Patient arrivals According to the CHE admission scheme and resource allocation, 4 type I and II patients are scheduled to undergo surgery during the allocated OR sessions, respectively. Type III patients arrive at the IC according a Poisson process with on average two patients per day. Type IV patient arrivals vary between 2 and 4 patients per day with a mode of 3. Their arrival is determined determined by the corresponding OR scheme, as explained in Section 2.3.5. As there is currently no coordination among the surgical specialties at the CHE concerning their OR schemes and resulting need for postoperative ICU beds, their arrival can also be considered as a random process. In discussion with CHE domain experts the obtained parameters and models were deemed realistic.

2.3.6 Technical details of implementation

The agent-based simulation model is implemented in Java as an event-based simulation [78]. In our model, events are patient admissions, transfers and discharges. The different local events at the care units trigger the communication among the involved agents through messages. The simulation runs on a single thread with the agent communication being synchronized as follows: the hospital simulator provides at every time stamp, i.e. every ten minutes in simulated time, the opportunity to the agents to send and respond to messages during multiple rounds. The simulation clock is moved to the next time stamp if no events have occurred or there are no more message to exchange in the current round. For sending and responding to messages the agents are selected in a random order in each round.

The waiting lists for the different patient groups are generated at beginning of a simulation run. For the patients the patients' arrival dates and pathway, i.e. the required treatment steps (including complications) and the respective LoS, are sampled. The information disclosed to an agent is restricted to an event at the time a patient can be transferred or discharged. In the case of a possible patient transfer, the hospital unit to which the patient is to be transferred is indicated.

The computation of the random samples from the different distributions and the ordering of agents in the interaction can be initialized by a fixed

random seed in the simulation which allows for the repetition of simulation experiments.

A run of 52 simulated weeks using the agent-based simulation takes on average about 0.7 seconds on an Intel Pentium 4 2.4GHz machine with 2GB RAM. Including the possibilities of re-transferring patients, cf. Section 2.3.3, increases the runtime of the simulation by approximately factor 5 due to the increased communication efforts among the agents.

2.4 Experimental evaluation

Our goal was to develop a realistic simulation for patient admission and transfer scheduling for surgical and emergency patient flows. For this aim, we discussed the agent-based simulation described in Section 2.3 during frequent meetings with domain experts at the CHE. Moreover, we performed various simulation experiments to compare the performance of the agent-based simulation to the outcomes achieved by human planners at the CHE hospital. This section describes the conducted experiments. First, we describe the experimental setup used for the simulation experiments. Second, we present the results obtained for the basic setting where the resource allocation and agents' policy parameters were obtained from the CHE case study. Then, we consider what-if scenarios relating to allocation adjustments and different patient (re-)transfer policies employed by the care unit agents in the simulation.

2.4.1 Setup of simulation experiments

For validating and evaluating the simulation we performed 50 simulation runs of 52 simulated weeks, each after a warming-up period of 12 weeks. Preliminary experiments showed that a warming-up period of 12 days is sufficient in order to avoid starting with an empty hospital.

Patient pathway settings The patient pathway settings of our simulation experiments are given in Table 2.3 on page 50 which are based on the CHE data analysis. We employed the admission scheme of the CHE as described in Section 2.3.5. The number of type III and IV admissions is restricted by the bed availability at the corresponding units.

Patient inflow at the MC is included in an abstract manner: the number of available beds is sampled at the start of a day using a discrete stationary probability distribution. This representation was chosen because type I

and II patients are admitted to the MC for about one day (on average). This implies a minimal time dependency between subsequent days. Other patient inflow requires 3, 2, 1 or 0 beds with probability 0.2, 0.5, 0.2 and 0.1, respectively.

Resource allocation settings For the basic validation and evaluation of our system, we consider the current fixed resource allocation employed at the CHE given in Table 2.1 on page 45. The associated relative unit costs, $c_u\, u \in U$, are given in Table 2.4. The unit costs for the different types of hospital beds relate to the daily costs for staff and materials and are expressed relative to the costs of a nursing ward bed. For example, at the IC the ratio between patients and nursing staff is 1:1 whereas the ratio is 2:1 at the MC.

Definition and calculation outcome measures As reported in Section 2.3.6, the simulation model offers a number of outcome measures. Of particular interest to the hospital management is the patient throughput defined as the number of patients that leave the system after completed treatment, for the different patient groups.

Hospital resource costs refer to "regular" resources and their associated costs based on the resource allocation at the hospital units. Specifically, the resource costs for "regular"'resource capacity are calculated by the sum of the number of allocated resources, r_u weighed by the respective cost parameters, c_u, for $u \in U$. The total resource costs are given by the sum of the costs for the "regular" resource capacity and the costs arising from unused OR capacity, i.e.

$$\text{total resource costs} = \sum_{u \in U \setminus \{CTS-OR\}} c_u \cdot r_u + c_{CTS-OR} \cdot uc_{CTS-OR}, \quad (2.2)$$

where uc_{CTS-OR} denotes the period of unused OR capacity due to canceled surgeries as a result of unavailable postoperative care beds. The cost factor c_{CTS-OR} comprises personnel costs for surgeons, anesthesiologists and other OR staff scheduled for the OR session. We assume that all fixed and variable costs for an OR are covered by the surgical procedure that is to be performed. Therefore, costs from unused OR capacity arise since a higher OR utilization would result in reduced staff requirements for the OR. The cost calculation does not include sunk costs for OR, hospital beds and equipment as capital budgeting purposes are beyond the scope of the developed simulation.

Unit	CTS-OR	CTS-PACU	CTS-HC	IC	IC-HC	MC	CTS-ward
$\mathbf{c_u}$	0.09	2	2	4	2	2	1

Table 2.4: Relative unit resource costs in the CHE case study

Back-up capacity usage is expressed as the total sum of time, i.e. days, that back-up capacity was used at the different hospital care units weighted by the respective unit's resource costs.

2.4.2 Basic scenario

In Table 2.5 the mean and standard deviation (stdev) of the outcome measures for the setup described in Section 2.4.1 with the current CHE resource allocation given in Table 2.1 are shown. The results presented in this section do not involve the re-transfer of alternatively accommodated patients to the units clinically indicated by their respective pathway.

Outcome measure	**mean ± stdev**
Throughput	
type I+II	1844.84 ± 30.61
type III	540.8 ± 23.59
type IV	905.78 ± 8.75
Costs	
resource allocation	111.0 ± 0
canceled CTS surgeries	21.23 ± 2.74
Total back-up capacity usage	428.04 ± 54.55

Table 2.5: Simulation outcomes (mean ± standard deviation) for basic scenario

With the decision policies presented in Section 2.3.3, the agent-based simulation achieved a mean total patient throughput of about 3300 patients. Of this, 1845 patients of type I and II are treated with a standard deviation of approximately 30. Purely based on the CTS-OR capacity, a maximum throughput of 2080 type I and II patients could be realized. In practice, however, this upper bound is not realized because the frequent blocking at the ICU affects the patient flow through CTS-PACU and CTS-HC which in turn causes canceled CTS surgeries. At the CHE, about 1800 type I and II patients undergo surgery per year. Thus, the output of the agent-based simulation compares well to the human CHE planners. Regarding admission requests for type III and IV patients, the system achieves an acceptance rate of about 84% and 99%, respectively. These outcomes are comparable to recent aggregated measurements performed at the CHE.

In our simulations, about 430 days of back-up capacity are used in one year, cf. Table 2.5. This corresponds to approximately 1.2 beds per day. The majority of the back-up capacity usage (95%) is attributed to the CTS-ward, the remaining 3% and 2% of back-up capacity is used by the IC and the MC unit, respectively. At the CTS-ward 54% of the back-up capacity is required to admit preoperative patients after which these patients follow the regular treatment process described in Section 2.3.5. On average, back-up capacity is used for about 14 hours per day. Postoperative patients are initially admitted to a back-up bed in about 15.5% of the cases with a mean LoS of about 7.5 hours. At the IC a back-up bed is required about once every two weeks for circa 9 hours on average. The frequency of back-up capacity usage at the MC back-up capacity is slightly higher compared to the IC (once every ten days), in total for about 7 days per year. Domain experts from the CHE have found the above results to be realistic.

Simulated patient flows Due to the alternative patient routing policies employed by the care unit agents, the simulated patient routing differs slightly from the routing indicated by the corresponding patient pathway parameters, cf. Section 2.3.3. The simulated patient routing on which the results in Table 2.5 are based is given in Table 2.6. The flow of type I patients deviates from the pathway routing such that about 4% of the patients with MC indication are alternatively admitted to the IC-HC or IC due to the unavailability of MC beds. This corresponds well to the historical flow of type I patients at the CHE found in our data analysis, cf. Section 2.3.5. The type II patient flow closely follows the clinical pathway as can be expected since alternative routing policies do not involve type II patients, cf. Table 2.2. Thus, compared to the CHE case, the simulation not only achieves comparable overall outcome measures but also closely resembles historical patient flows for type I and II patients.

For type IV patients, solely 62,5% of the patients are admitted to the unit indicated by the patient pathway. The remaining 37,5% of the patients are in almost equal shares admitted to the IC and the CTS-HC unit. Unfortunately, the conformance of type IV patient routing could not be evaluated using CHE case study data. However, since the patient throughput results in Table 2.5 correspond to aggregated CHE measurements the simulation appears to approximate the real-life routing well.

Simulation run length The simulation outcomes are almost linear in the number of simulated weeks in a simulation run, as shown in Figure 2.8.

Patient group	Unit	Routing param. patient pathway	Simulated routing prob.
Type I	IC	0.15	0.153
	IC-HC	0.0	0.034
	MC	0.15	0.112
	CTS-ward	0.7	0.701
Type II	IC	0.05	0.052
	MC	0.15	0.147
	CTS-ward	0.8	0.801
Type IV	IC-HC	1.0	0.625
	IC	0.0	0.189
	CTS-HC	0.0	0.186

Table 2.6: Pathway routing parameters, cf. Table 2.3, and simulated patient routing resulting from care unit agents' scheduling policies

The total throughput ranges between about 253±9 to more than 3300±40 patients. The values of the other performance measures have a comparatively smaller range with about 113±3 to 133±12 and 30±11 to 453±50 for the total resource costs and the back-up usage, respectively. Therefore, for computationally expensive calculations also shorter durations of the simulation can be used, e.g. for resource occupancy prediction and allocation optimization, presented in the subsequent chapters of this thesis.+

2.4.3 Scenario analyses

In order to better understand the complex relationship between resource allocation, scheduling policies, resource occupancy and admission acceptance rates, we analyzed several scenarios using the agent-based simulation model described above. Here, we present multiple parameter settings relating to the decision variables in the simulation model. First, we consider changes in the allocation of resources at the different hospital units and their effect on the system's performance. We examine the allocation of CTS-ward beds in order to decrease the local back-up capacity usage and analyze the resulting duration and frequency of back-up capacity usage. Moreover, we consider the scenario where the IC-HC is closed and the allocated beds are shifted to the IC to increase the flexibility of resource usage at the ICU. Second, we address the impact of the scheduling policies of the agents on the system performance. Here, we consider the effect of re-transfer policies employed by the IC, IC-HC and CTS-HC agents on the resulting patient throughput, resource costs and back-up capacity usage. Furthermore, we present

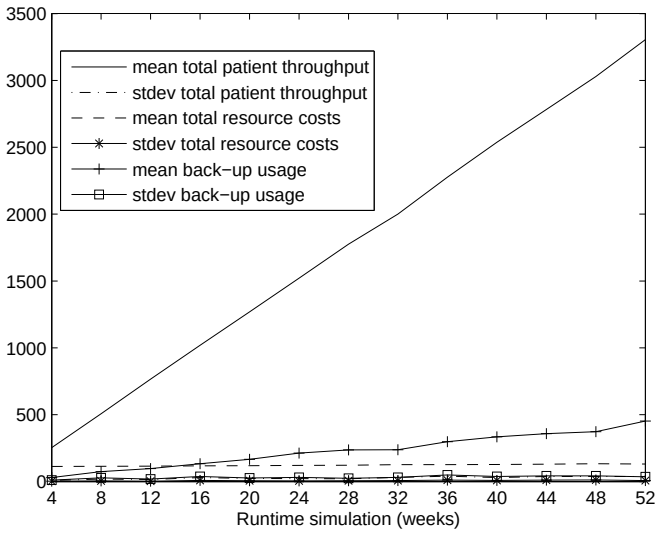


Figure 2.8: Simulation outcomes as a function of the duration of a simulation run including 12 weeks of warming up

an adaptive method for employing re-transfer policies at the different care units based on the current resource utilization.

Resource allocation at CTS-ward

In the current situation at the CHE, cf. Section 2.4.2, about half of the preoperative CTS patients are admitted to (back-up beds at) other nursing wards because no bed is available at the CTS-ward. Although the quality of care is not compromised, admission to the CTS-ward is preferable from a patient-friendliness point of view[3].

In order to reduce the back-up capacity usage at the CTS-ward, we evaluated the resulting back-up capacity usage from varying the number of CTS-ward beds between 35 (the current CHE allocation) to 42 beds. Preliminary experiments showed that more than 42 CTS-ward beds did not result in further reduction of back-up capacity usage. The resource allocations at the other hospital units remain unchanged. Tn Figure 2.9 the mean and standard deviation (depicted by the bars) of the duration of the back-up capacity usage (in days) and the frequency for 50 simulation runs of 52 weeks each are depicted. It should be noted that the CTS-

[3]Admission to the CTS-ward is preferred as it familiarizes the patient with the unit and staff and facilitates additional tests and consults if required.

ward capacity does not restrict patient admissions and transfers, thus the throughput remains constant at about 1844 patients p.a.

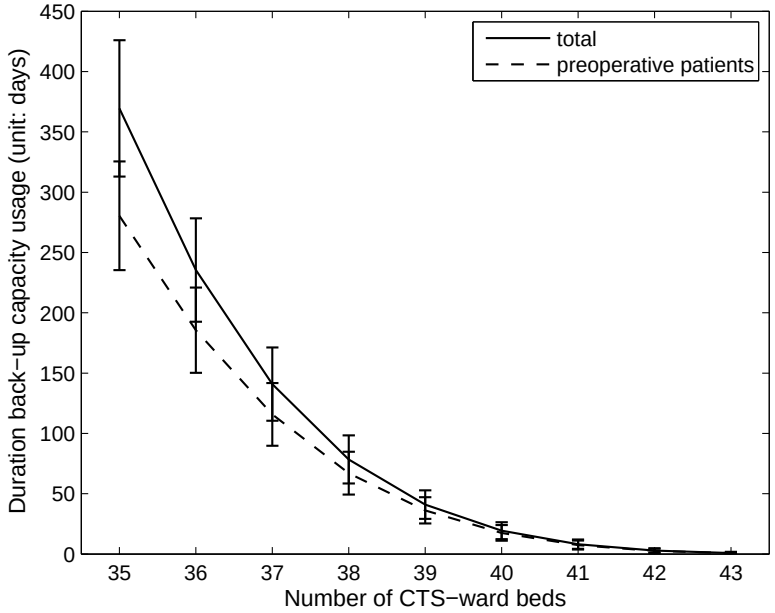
From a cost perspective, additional CTS-ward beds linearly increase the total resource costs according to equation (2.2). With respect to the usage of back-up capacity, we can see that the number of CTS-ward beds is inversely proportional to the duration and frequency of back-up capacity usage at the CTS-ward, as was expected. The curve progression in both figures suggest a hyperbolic relationship. The marginal reduction in bach-up capacity usage duration and frequency decreases for increasing number of CTS-ward beds. For 36 CTS-ward beds the back-up capacity usage decreases by about 35%. For a higher number of CTS-ward beds the marginal decrease in duration and frequency becomes smaller with about 26%, 16% and 10% for 37, 38, 39 and 40 CTS-ward beds, respectively. An absolute reduction in back-up capacity usage of 80% and 90% is achieved for 3 and 4 additional CTS-ward beds.

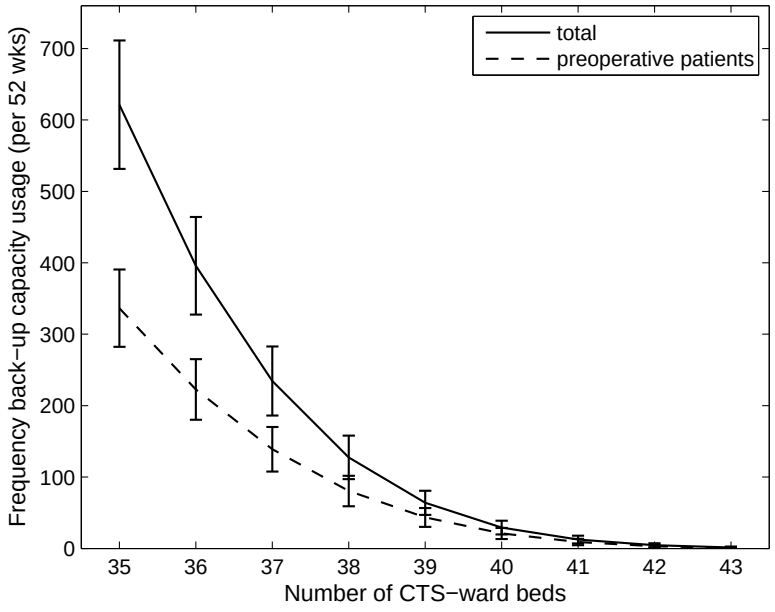
Back-up capacity is chiefly required for accommodating preoperative patients of which the share increases for increasing number of CTS-ward beds. Moreover, we can note that the standard deviation of back-up capacity usage decreases with increasing CTS-ward capacity.

Intermediate conclusions For a resource allocation decision, a trade-off has to be made by hospital management between feasible back-up usage and resource costs. The marginal reduction in back-up usage, both in terms of the duration as the frequency, diminishes for increasing CTS-ward capacity.

Resource allocation at IC and IC-HC

In favor of increased flexibility in patient admissions at the ICU, the CHE management discussed the option of closing the IC-HC unit and transfer the bed capacity to the IC. Using the simulation model various allocations of beds at the IC can be analyzed. The resource allocations at the other hospital units remain unchanged. Figure 2.10 shows the mean and standard deviation of the realized throughput per patient group for the scenario with 0 IC-HC beds and varying IC bed capacity. For increasing IC bed capacity, the figure demonstrates that the mean throughput of CTS, other surgical and emergency patients increases with decreasing variability. Specifically, the mean throughput of type I and II patients increases almost linearly from 1562.3 to 1975.92 for 10 to 17 IC beds. For more than 17 IC beds the increase diminishes with in total about 2051 type I+II patients being treated if 20 beds are allocated to the IC. Interestingly, the standard deviation increases





(b) frequency

Figure 2.9: Mean and standard deviation (bars) of back-up capacity usage at CTS-ward for varying number of CTS-ward beds in terms of (a) the duration in days and (b) the frequency

for increasing number of IC beds ≤ 13 with about 46% from 30.6 to 44.6 and then decreases for more than 14 beds by about 61% to 17.3. It appears that an allocation with a small number of IC beds and no IC-HC beds causes an increased interference between the different patient flows in the system. The throughput of type III patients is on average about 500 with a standard deviation of circa 23 for 10 IC beds and increases to a mean of about 630 with a standard deviation of about 19 for 17 IC beds which corresponds to an acceptance rate of approximately 98%. The mean and standard deviation of the type III throughput remain almost constant for more than 17 IC beds. Thus, the variability in type III throughput is primarily determined by the variation of patient arrivals and LoS. The type IV throughput remains almost constant with a mean increase of about 30 from 10 to 20 IC beds which corresponds to an acceptance rate for admission of 99.7%. Here, the increasing number of IC beds causes the the number of type IV patients treated at the CTS-HC to decrease and to increase the number of treated type IV patients at the IC. Thus, we can conclude that the patient mix is strongly dependent on the resource allocation in the system. In order to guarantee the same total patient throughput as in the CHE case study 3 additional IC beds are required when the IC-HC unit is closed. A total of 15 IC beds is needed in order to achieve a patient mix with a minimal throughput per group that is comparable to the CHE patient mix if the IC-HC is closed. Thus, the resource allocation at the ICU has a great impact on the patient mix flowing through the system and on the variability of the patient flows.

Figure 2.11 depicts the total resource costs and the total back-up capacity usage in the network of care units for different resource allocations with 0 IC-HC beds and varying number of IC beds. The costs are a convex function of the number of IC beds and range from 152.82 to 144.53 for 10 to 20 IC beds with a minimum at 16 IC beds and costs of about 140.7 thereby exceeding the costs for the CHE resource allocation of about 132.45, cf. Table 2.5. Thus, from a cost perspective the closing of the IC-HC appears to be disadvantageous compared to the resource allocation in Table 2.1.

Regarding the back-up capacity usage, closing the IC-HC affects the IC and the network of care units in different ways. On the one hand, the back-up capacity usage at the IC decreases from 10.9 to 0.2 bed days on average from 11 to 20 IC beds. On the other hand, the total back-up capacity usage in the network is an increasing function of the number of IC beds in this scenario ranging between about 158 to 788 beds days in total. For the allocation with 16 IC and 0 IC-HC beds that minimizes the resource costs, the back-up capacity usage amounts to about 559 beds days which is

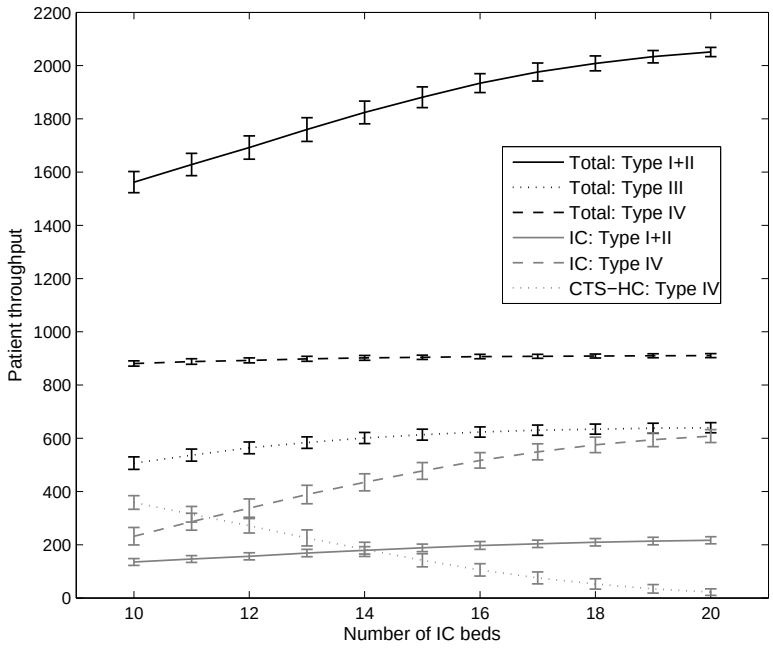


Figure 2.10: Mean and standard deviation (bars) of patient throughput realized with an allocation of 0 IC-HC beds and varying number of IC bed

about 30% higher compared to the current CHE allocation. This effect can be explained by the overall increased patient flow which is not aligned with the current resource allocation at the remaining care units. For minimizing the back-up capacity usage one should therefore consider not solely the IC and IC-HC unit, but all of the involved units in order to avoid creating mismatches between demand for care and resource availability.

Optimization of bed allocation at IC and IC-HC To automatically find an optimal bed allocation at the ICU, we implemented a brute-force optimizer that uses the simulation to evaluate different bed allocations. It can be used for various objective functions. The number of IC-HC and IC beds are varied from 0 to 15 and from 5 to 25 which results in 336 possible bed allocations. Each allocation was evaluated on the basis of 20 simulation runs of 52 simulated weeks and a warming-up period of 12 weeks. On an Intel Pentium 4 2.4GHz machine with 2GB RAM the runtime of the allocation optimizer amounts to about 81 minutes on average. We illustrate the algorithm using the cost-effectiveness ratio of the resource allocations as a one-dimensional objective function to be minimized. The cost-effectiveness measure is defined as the ratio of (1) mean resource costs (including the costs

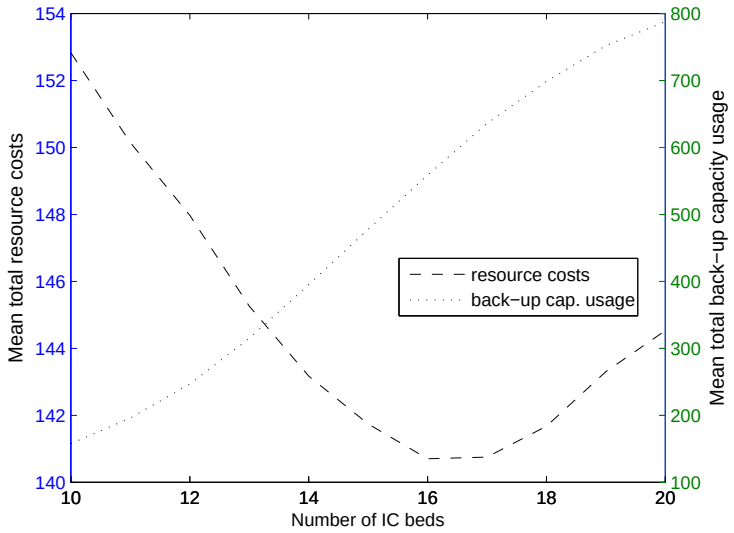


Figure 2.11: Mean total resource costs and back-up capacity usage realized with an allocation of 0 IC-HC beds and varying number of IC bed

for canceled surgeries) to the mean total patient throughput and (2) mean total back-up usage to the mean total patient throughput. In the first case the optimal bed allocation is 6 IC-HC beds and 5 IC beds which results in resource costs per patient of 0.033 cost units on average. The mean annual total throughput is about 3100 patients. In the second case an optimum is reached for 8 IC and 0 IC-HC beds with a mean back-up usage of 0.048 bed days per patient.

Figure 2.12 and Figure 2.13 show a plot of the corresponding objective landscape of the mean resource costs and back-up capacity usage per patient for different IC-HC and IC bed allocations, respectively. Figure 2.12 shows that resource costs per patient are convex with a minimum at 5 IC and 6 IC-HC beds. Compared to the current situation at the CHE, this allocation increases the patient throughput of type I+II by 4.55%. Type III throughput decreases by factor 2, while type IV throughput remains almost the same. Costs for regular capacity are decreased by 21%, whereas back-up costs are increased by 92.44%. For the mean back-up usage per patient depicted in Figure 2.13, mean costs increase by almost 39%, type I+II, III and IV throughput is decreased by about 34.4%, 16.2% and 18.5%, respectively. Back-up usage, however, is decreased with 84.3% compared to the performance of the current CHE allocation. These results show the

complex relationship between the patient throughput and the allocated resources in a situation with intersecting patient flows. Unlike from a costing perspective, decreasing the IC-HC capacity seems advisable from a back-up capacity usage point of view.

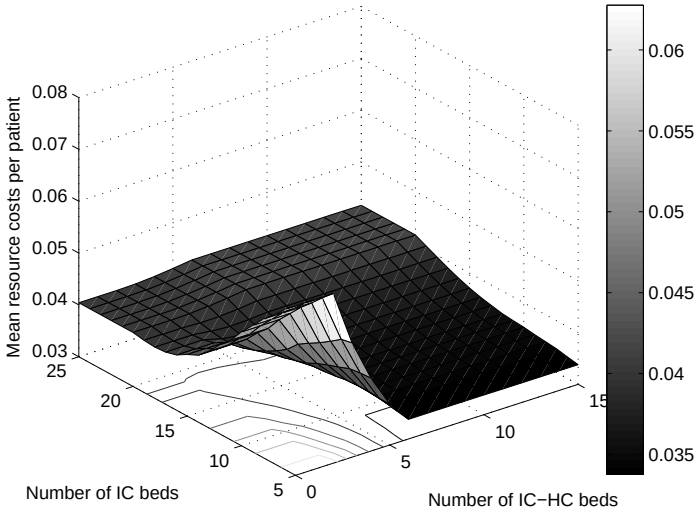


Figure 2.12: Surface plot of mean resource costs per patient for varying IC-HC and IC bed allocations

Intermediate conclusions Closing the IC-HC increases the proportion of treated type III patients and decreases the throughput of type I+II and IV patients. A resource allocation that yields a comparable patient mix to the CHE patient mix incurs higher resource costs but lowers the back-up capacity usage in the system. Thus, a trade-off between the patient throughput, the resource costs and the back-up capacity usage is required for allocation decisions concerning the IC. Moreover, we can conclude that the patient mix is strongly dependent on the resource allocation in the system. Therefore, the scope for capacity adjustments should comprise all the units in the network in order to avoid creating mismatches between demand for care and resource availability.

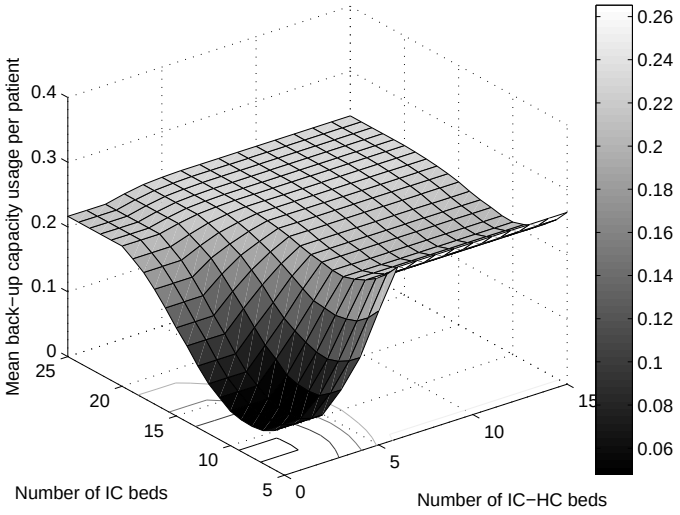


Figure 2.13: Surface plot of mean back-up capacity usage per patient for varying IC-HC and IC bed allocations

Effect of employing patient re-transfer policies

Another complex issue in hospital patient flow scheduling practice is the decision whether or not patients admitted to a unit that is not clinically indicated should be re-transferred to follow the originally provided patient pathway. For example, type IV patients could be admitted to a free CTS-HC bed in the situation of a temporary bed shortage at the IC-HC and be re-transferred to the indicated IC-HC as soon as a bed becomes available. From an organizational and patient perspective, patient re-transfers are undesirable as the transfers introduce additional disturbance in the patients' revalidation and nursing staff's working process. From an efficiency point of view, however, the possibility of re-transfers provides for the appropriate accommodation of patients after periods of (local) resource scarcity while maintaining patient flow and restoring patient pathways in the system.

The policy for re-transferring patients works as follows: the agent representing the care unit the patient was alternatively admitted to proposes regularly, e.g. every hour, the patient's transfer to the originally indicated unit such that the patient will be transferred as soon as a bed becomes available. A summary of the considered re-transfer destinations for the different patient groups is given in Table 2.7, cf. Table 2.2. As explained in Section 2.3.3, only type I patients with MC indication are considered for

re-transfer as type I patients with IC indication will always be admitted to the IC. Here, we consider the alternative routing policy for re-transfer of type I patients. In the following, we study several situations. First, we analyze the impact of fixed re-transfer policies at the different units. Then, we study an adaptive mechanism to regulate patient re-transfers depending on the capacity utilization at the different units.

	Patient groups			
Hospital unit	**type I**	**type II**	**type III**	**type IV**
CTS-HC	(1)MC, (2)IC-HC, (3)IC	-	-	IC-HC
IC-HC	MC	-	-	-
IC	-	-	-	IC-HC

Table 2.7: Possible re-transfer destinations (and ordering) for different patient groups at different care units

Effect of employing re-transfer policies at CTS-HC, IC-HC and IC
The results from applying re-transfer policies at the IC, IC-HC and CTS-HC for the different patient groups with the current CHE resource allocation are given in Table 2.8 on page 66. Here, we consider the diverse individual and joint options for re-transfer policies for patient groups and care units. The option of re-transferring type I patients from the IC-HC unit is omitted in Table 2.8 as the results almost equal the results of the 'type I+IV all units' option.

The different options for applying re-transfer policies have varying effects. Solely considering type I patients for re-transfer yields same or slightly increased patient throughput and back-up capacity usage compared to the basic setting without re-transferring. Re-transfer policies that relate to type I and type IV patients, however, reveal throughput improvements of considerable extent for type I, II and III patients, minor changes for type IV patients and considerable reduction of canceled CTS surgeries. Similar to the resource allocation scenarios above, the increased patient throughput comes at the costs of increased back-up capacity usage. Re-transferring all type I and IV patients at IC, IC-HC and CTS-HC yields the best cost-efficiency ratio, followed by the option to re-transfer all type I and IV patients at IC and CTS-HC.

Outcome measure	Re-transfer policy			
	type I+IV all units	type I all units	type IV IC	type I+IV CTS-HC
Throughput				
type I+II	1925.56 ± 25.72	1850.12 ± 30.98	1866.6 ± 36.63	1905.82 ± 23.72
type III	555.4 ± 24.11	540.3 ± 23.86	558.26 ± 24.53	538.68 ± 23.53
type IV	910.38 ± 7.86	906.68 ± 8.62	906.68 ± 8.13	910.14 ± 7.94
Costs canceled CTS surgeries	14.09 ± 2.21	20.98 ± 2.78	19.48 ± 3.25	15.91 ± 2.05
Total back-up capacity usage	531.44 ± 58.58	436.46 ± 58.09	459.89 ± 63.45	494.84 ± 52.93
	type I CTS-HC	type I+IV IC & CTS-HC	type I+IV IC & IC-HC	type IV all units
Throughput				
type I+II	1844.86 ± 31.39	1922.52 ± 26.87	1921.86 ± 27.41	1871.42 ± 38.18
type III	540.4 ± 23.82	555.6 ± 24.14	555.34 ± 24.22	557.76 ± 24.54
type IV	906.02 ± 8.65	910.24 ± 7.83	909.72 ± 7.96	907.08 ± 8.25
Costs canceled CTS surgeries	21.45 ± 2.8	14.33 ± 2.3	14.4 ± 2.35	19.033 ± 3.41
Total back-up capacity usage	428.78 ± 55.24	525.16 ± 56.09	525.61 ± 57.55	469.03 ± 64.23

Table 2.8: Simulation outcomes (mean ± stdev) for different combinations of re-transfer policies employed by IC, IC-HC and CTS-HC

Adaptive re-transfer mechanism Due to the beneficial performance of re-transferring patients we also considered an adaptive mechanism that controls whether re-transfer policies are employed when a high degree of resource utilization is achieved. The resource utilization is defined as the state at unit u at time t, denoted by $s_u(t)$, which is given by the ratio between utilized capacity and allocated resources, i.e.

$$s_u(t) = \frac{\text{no. beds occupied at } u \text{ at time } t}{r_u}. \tag{2.3}$$

The mechanism is based on a predefined threshold for the resource utilization at unit u, $\mathcal{RUT}_u$, and applies to the specified patient types I and IV, where applicable. The policy can be described as follows:

$$\begin{cases} \text{propose pat. temporarily admitted to } u \text{ for transfer} & \text{, if } s_u(t_i) > \mathcal{RUT}_u, \\ \text{pat. remain admitted to } u \text{ until } LoS_u^g \text{ has elapsed} & \text{, otherwise.} \end{cases} \tag{2.4}$$

So, if a unit is too crowded, i.e. the threshold of resource utilization is exceeded, the respective care unit agent proposes patient re-transfers to the agents given in Table 2.7 on page 65. Otherwise, already admitted patients remain at the respective unit until the patients' LoS has elapsed and they are eligible for normal transfer.

Specifically, the resource utilization is evaluated on a hourly basis and the re-transfer policy is set accordingly. The simulation outcomes realized by the adaptive mechanism for different utilization thresholds are given in Table 2.9. The actual patient routing that results from the patient re-transfers is comparable to the routing presented in Table 2.6.

The utilization thresholds in Table 2.9 are based on utilization targets typically presented in the literature, e.g. Vissers and Beech [96]. In a sensitivity analysis we also considered thresholds below 0.7 and in steps of 0.05. However, these settings resulted in comparable results and are therefore omitted in Table 2.9. The adaptive mechanism can achieve a slight improvement compared to the performance of the fixed re-transfer policies in Table 2.8. A threshold of 0.8, for example, realizes approximately the same mean patient throughput, CTS surgery cancelations and back-up capacity usage but reduces the variability of the outcome measures. Higher threshold values provide for slightly decreased patient throughput accompanied by reduced back-up capacity usage and increased surgery cancelations.

Intermediate conclusions Employing re-transfer policies can considerably increase improve the patient flow in the simulation. A further improvement in efficiency can be achieved through adaptive re-transfer policies that

	Utilization thresholds $\mathcal{RUT}_u$			
Outcome measure	**0.7**	**0.8**	**0.9**	**1.0**
Throughput				
type I+II	1925.54 ± 25.71	1925.34 ± 24.4	1922.86 ± 25.39	1915.8 ± 25.13
type III	555.4 ± 24.11	555.74 ± 23.81	554.98 ± 23.64	555.14 ± 24.11
type IV	910.38 ± 7.86	910.38 ± 8.04	910.18 ± 8.13	910.14 ± 7.99
Costs canceled CTS surgeries	14.09 ± 2.21	14.09 ± 2.09	14.31 ± 2.18	15.0 ± 2.16
Total back-up capacity usage	530.39 ± 58.29	526.44 ± 56.62	519.02 ± 57.68	507.12 ± 51.99

Table 2.9: Simulation outcomes (mean ± stdev) for adaptive re-transfer mechanism with varying utilization thresholds

initiate patient re-transfers only if the resource utilization rate exceeds a predetermined threshold. Selecting an appropriate threshold involves a trade-off between patient throughput and resulting back-up capacity usage.

2.5 Conclusions

In this chapter we presented an agent-based simulation for hospital patient scheduling that realistically captures the complex features of the problem domain. To the best of our knowledge, this is the first agent-based simulation for patient admission and transfer scheduling that includes multiple patient groups with stochastic arrival and treatment processes. We showed that an agent-based simulation can be developed based on knowledge elicitation from the case that realistically reflects the problem domain. Extensive simulation experiments demonstrate the applicability of the simulation and show how the agent-based simulation is useful for decision support. Furthermore, the implemented simulation can be adjusted to comparable situations in other hospital settings. In a hospital setting where the planning is often performed in a decentralized way, a multi-agent approach is ideal because it allows for designing and evaluating improved (adaptive) policies, which can then be implemented easily in real life.

The multiple simulation outcomes for the basic setting show that performance achieved by the agent-based simulation is comparable to the planning performed by hospital staff of the CHE. The patient throughput differs minimally from the situation at the case study hospital and CHE planners and managers consider the results on resource costs and resource usage realistic for the basic scenario.

What-if scenarios show that the agent-based simulation can be helpful in analyzing the complex relationship between bed allocations, occupancy and patient mix. The agent-based simulation approach allows for a detailed modeling of the decision making processes of the involved care units and realistic analysis of the system that otherwise would be impossible. Moreover, the simulation allows for a fast analysis of changed input settings where one year of hospital time can be simulated in less than one second. Thus, the simulation is of substantial value for decision support for hospital management.

We also presented a first approach to optimize the resource allocation in the network of care units using the simulation. Here, we considered the number of IC-HC and IC beds as free variables which appeared to have a significant influence on the overall patient throughput. The efficient computa-

tion and the size of the search space allowed using a brute-force optimization which guarantees a globally optimal solution. We illustrate the optimizer by using the mean resource costs and mean back-up capacity usage per patient as objective functions, but other performance measures can also be easily incorporated in the simulation. Our results demonstrate that a local allocation decision affects the patient flows not only at the respective care unit, but in the entire network of care units in the case of the complex patient pathways considered in this thesis. Therefore, allocation decisions should be taken at the level of the network of care units and thus coordinate all relevant resource categories to avoid creating mismatches between demand for care and the resource availability. Also, our results show the multi-objective nature of the hospital resource management problem. These two aspects will be further addressed in Chapter 4 and extended to allow for dynamic allocations in Chapter 5. The complex stochastic problem features, the decentralized and realistic decision making, the short runtime of the simulation and the insightful results presented in this chapter show that a well-designed agent-based simulation for hospital scheduling is of substantial use for optimizing resource allocation decisions in this complex problem setting. It allows the optimized allocation policies to be implemented easily in real life.

Moreover, we addressed different settings for the decision making process of the different agents. Specifically, we considered employing re-transfer policies to schedule patients that were temporarily admitted to a care unit that is not clinically specified to the originally indicated unit in order to reduce possible interference between patient flows. We showed that a considerable improvement in patient throughput can be achieved by an unchanged resource allocation by re-transferring both type I and IV patients and proposed a first adaptive mechanism which reduced back-up capacity usage and the variability of the simulation results. In the following chapters, however, we will not consider re-transfer policies employed by the care unit agents in order not to optimize resource allocation or admission schedules that lead to frequent patient re-transfers. Re-transfer policies should rather be implemented aiming at increasing operational flexibility for patient accommodation in situations of local resource shortage.

An additional advantage of the agent-based simulation presented in this chapter in comparison with other, for example mathematical, modeling approaches is its flexibility. The implemented model for patient pathways allows for an uncomplicated adjustment of the simulated pathways through adjusting the respective pathway parameters. In a mathematical model changing the structure of the modeled processes is not straightforward. Moreover, the detailed and modular level of modeling facilitates the adjust-

ment of the simulation to particular hospital settings through adjusting the scheduling rules, the number and type of agents and agents' decision policies. This detailed representation of a specific hospital setting would not be as straightforward using a simple discrete-event simulation approach. Furthermore, the generic setup of the simulation allows to easily incorporate more complex medical decision making policies for selecting patients in periods of resource scarcity.

An interesting extension to the analysis presented in this chapter would be to explore possibilities for hospital management to steer the patient mix through the resource allocation decision. In our experiments, we studied the effect of different allocations on the resulting patient mix. In a first attempt to control the patient mix, the presented optimization approach could be applied to evaluate a large range of possible resource allocations and determine the resource configuration(s) that realize(s) a predetermined patient mix.

A limitation of the simulation in its current implementation are the static waiting lists for elective surgery which are assumed to be sufficiently long, so elective patients are always available for admission. This assumption is valid for the Netherlands and several other European countries where the waiting list for cardiac surgery are long. Moreover, specialists can keep a relatively constant number of patients on the waiting lists by either changing the criteria for admission or by referring patients to other specialists or hospitals which are both used in practice Groot [37].

Given the practical case study, the validation of our model is a complex issue. The current practice and historical data provide only a single instance, and it is difficult to identify appropriate performance indicators for a wide range of settings. Additional to historical data, to which the patient throughput and flow achieved by the simulation corresponded well, we evaluated the simulation and its elements in numerous meetings with domain experts and planners at the CHE. In our discussions the agent-based simulation and the obtained results were well received. Because of the realistic modeling and the promising results, the simulation will be used at the CHE for further analysis and optimization of patient flow control.

Chapter 3

Prediction of hospital resource usage

In this chapter we focus on predicting future resource usage based on current occupancy information and planned patient admissions. We evaluate two approaches for resource-usage prediction: forward simulation using the simulation described in Chapter 2 and supervised learning using neural networks. We assess the two approaches with respect to a benchmark prediction heuristic derived from the hospital case study. Moreover, we analyze the underlying resource-usage probability distributions and the applicability of different prediction statistics to be used for decision support. The first approach will be used further throughout the thesis. This chapter has contributed to the publication of Chapter 5 which will appear as [49].

3.1 Introduction

Prediction is concerned with the estimation of future and unknown events based on current and past observations. Specifically, our goal is to predict the resource usage at a hospital unit one or more days in advance. Prediction is especially important for decision-making support in patient flow logistics because of its highly dynamic and time-dependent character, i.e. decisions taken now may influence the possible decisions to be taken in the future. Here, today's patient admission, transfer and resource allocation decisions influence the current and future resource usage at the different units in the hospital which in turn restricts possible future decisions. Prediction of future hospital resource usage can assist resource management and patient admission control in order to improve hospital operations. With respect to

resource management, where allocations are based on past experience, predicting future resource need can help deploying resources more efficiently in order to meet the current and future demand. Regarding patient admission control, predicting the effect of an admission decision on the resource occupancy can help anticipate future bottlenecks that possibly require back-up capacity usage or cause the blocking of patient flow and increase patient waiting times. Therefore, prediction of future resource usage is important for assisting resource management and patient admission planning in hospitals.

Predicting the future resource usage in a hospital setting is a complex problem, especially since the resource usage at a hospital unit is highly dynamic and uncertain. The fluctuations in resource occupancy are due to patient admissions and transfers occurring continuously over time, patients' length of stays being stochastic and mostly unknown beforehand, unexpected patient transfers caused by complications and emergency patients arriving in urgent need for care, cf. Chapter 2. Often, resources like at the ICU are shared by different types of patients each with different resource needs for their treatment processes. Moreover, patient pathways often involve several hospital units that need to be taken into account.

In this chapter we evaluate techniques for predicting the future resource usage given the current resource occupancy and the planned patient admissions at the prediction moment. We consider two settings for determining future bed occupancy: (1) unconstrained admission control where patient admissions are mainly determined by the actual demand for care and the available resource capacity and (2) constrained admission control where a predefined admission scheme imposes an additional constraint on the maximal number of patient admissions, cf. Chapter 1. The resource occupancy is modeled as a probability distribution to be predicted which enables to represent numerous possible realization scenarios. Moreover, our modeling approach is very flexible as it allows derive manifold descriptive statistics to be used for decision support in a hospital logistics setting. The prediction approaches we evaluated are forward simulation and supervised learning using artificial neural networks (ANNs). Forward simulation involves the simulation of future time intervals given the current situation for a number of scenarios. Using a validated simulation as is the case here, forward simulation can provide precise predictions of the future at the expense of increased simulation time. In order to reduce the computational effort associated with forward simulation we analyze whether supervised learning can provide sufficiently accurate predictions. ANNs have been chosen as supervised learning technique as they are able to learn arbitrary dependencies

from measured data (in the presence of noise) provided infinite complexity of the network structure [7]. Greater complexity of an ANN, however, also entails an increase of the computing time. We study three widely-applied network structures to evaluate the ANNs' predictive power for the resource-usage prediction problem and compare the obtained predictions to a benchmark heuristic derived from the hospital case study. Moreover, we extensively analyze the predicted resource-occupancy distributions in terms of their shape, location and variability and discuss appropriate predictive statistics and distribution approximations. In our analysis, we present the obtained predictions in an explorative fashion. To the best of our knowledge, this is the first forecasting approach using simulation and distributions for hospital resource occupancy prediction.

The remainder of this chapter is organized as follows. First, we discuss related work on resource occupancy prediction in hospital settings in Section 3.2. Then, the underlying problem formulation and model are described in Section 3.3. Forward simulation is outlined and evaluated in Section 3.4, which is followed by a description of our supervised learning approach in Section 3.5. Finally, we provide our conclusions in Section 3.6. In the remainder the terms forecasting and prediction will be used interchangeably.

3.2 Related work

Earlier work on prediction models for hospital utilization have been presented both in the mathematical operations research and the computational intelligence literature. A mathematical prediction model for admission control is presented in Groot [37] that uses flow-based linear equations derived from simplified deterministic patient pathways. This approach is not applicable in the problem setting considered in this thesis that is characterized by complex stochastic patient pathways. The prediction approaches presented in this chapter account for the complexity and stochasticity present in the problem domain and use online occupancy information to model arbitrary non-linear dependencies. In Vissers [97] a long-term care demand model is presented which considers the resource usage resulting from demographic changes in the hospital's catchment area and is based on average LoS calculations. However, the model is not applicable here as decision support on resource management and admission control in our problem setting requires resource utilization predictions on the level of individual hospital care units. Our work can derive predictions both at care unit and hospital level and incorporates complex stochastic patient pathways. The work

in Kolesar [58] and Ridge et al. [80] propose queueing models for evaluating the resource usage at a single hospital unit. However, these models are not suitable for the setting considered in this thesis as the different units' resource occupancies are mutually dependent due to the complex patient pathways that involve multiple hospital units. Moreover, our prediction models are flexible and adjustable to other hospital and pathway settings. Also, they allow for evaluating short-term fluctuations of resource usage resulting from an admission scheme as well as the long-term behavior of the system. A time-series approach is developed in Tandberg and Qualls [91] for predicting patient arrivals, their acuity and the patients mean LoS at an emergency department. However, their results indicate that time-series forecasts of length of stay and patient acuity perform poorly and are not likely to contribute useful information for resource allocation decisions. Therefore, we consider more sophisticated forecast techniques as forward simulation and artificial neural networks in this chapter. The mathematical prediction model for patient admission control presented in Vissers et al. [99] applies to simplified deterministic patient treatment processes which is extended to probabilistic LoS in Adan et al. [1]. However, their model is restricted to patient pathways with deterministic routing and is therefore not applicable in the setting studied in this thesis. Moreover, with the focus being on tactical decision support for admission scheduling, their model is limited to the evaluation of the mean resource usage resulting from an admission decision. The approaches presented in this chapter consider the underlying resource-occupancy distribution and we investigate appropriate prediction measures that use information on the shape and dispersion of the distribution. In the analytical model presented in Kusters and Groot [60] a normally-distributed bed occupancy distribution is predicted based on the mean and variance of the current occupancy, the number of patient transfers and future (planned or uncertain) patient arrivals. However, the distribution may not be applicable to settings where the normality assumption does not hold. Our prediction approaches are distribution-free methods that also consider alternative patient transfers due to bed shortage at a hospital unit, which significantly influences resource usage in the system of hospital units considered in this thesis.

In the computational intelligence literature several neural network approaches have been presented, e.g. [61, 62, 102], to estimate patient LoS based on clinical variables. The proposed models use broad categories of LoS which limits the applicability of these prediction approaches for decision support in hospital patient flow logistics.

3.3 Model for admission control and occupancy prediction

In the following section we briefly recapitulate the model for admission control used in the agent-based simulation, described in detail in Chapter 2, and describe our model for prediction of resource usage at a hospital unit.

3.3.1 Admission control in agent-based simulation

As described in Section 1.1.3, we consider a discrete time period denoted by T, with equidistant prediction moments denoted by $t_i \in T$ with $t_{i-1} < t_i$ for $i = 1, 2, \ldots, n-1$. Formally, an admission scheme for the time interval $[t_i, t_{i+h}]$ is denoted by $\mathbf{a}_{[t_i,t_{i+h}]}$ and is given by $\mathbf{a}_{[t_i,t_{i+h}]} = (a^g_{t_k}, g \in \Theta, t_k \in [t_i, t_{i+h}])$ with $a^g_{t_i} \in \mathbb{N}_0$ referring to the maximum number of patients of group $g \in \Theta$ to be admitted to the hospital on day t_i. Let $\mathbf{a}_{t_i} = (a^g_{t_i}, g \in \Theta)$ be the admission scheme on day t_i specifying the (maximal) number of planned patient admissions for all groups of patients on day t_i.

For the OR scheduling and the care unit agents in the simulation described in Chapter 2 the introduction of an admission scheme imposes an upper limit on the possible number of patient admissions of the corresponding patient types with the actual admission decisions being determined on the basis of the scheduling policies presented in Table 2.2 on page 46. Thus, $a^{III(IC)}_{t_i}$ and $a^{III(MC)}_{t_i}$ affect the admission of type III patients on day t_i while $a^{I}_{t_i}$, $a^{II}_{t_i}$ and $a^{IV}_{t_i}$ limits the patient admissions of type I, II and IV patients on day t_{i+1} since a lead time of one day is employed in the simulation, cf. Section 2.3.5.

As explained in Chapter 1, Section 1.1.3, the actual admission of patients in the simulation depends not only on the respective admission scheme $a^g_{t_i}$, but also on the admission policy of the responsible agent (described in Section 2.3.3), the number of available resources at the respective unit and the demand for care of patient group g. Thus, $a^g_{t_i}$ can be interpreted as an upper bound for the number of actual admissions. This implies that an admission scheme imposes an additional constraint on the patient admissions depending on the value of $a^g_{t_i}$, i.e. the lower the value of $a^g_{t_i}$ the more influence $a^g_{t_i}$ has on the actual patient admissions compared with the other three control factors.

In the following we distinguish between two situations: unconstrained and constrained admission control. In the first case, the number of planned admissions is set to sufficiently high values such that the admission scheme imposes no additional constraint on the actual patient admissions. For the

constrained case, the admission schemes are determined online using a local search algorithm. The two problems are described in detail in Section 3.4.

3.3.2 Resource occupancy prediction

The approach presented in this chapter strives to predict the future resource occupancy on the basis of a constant resource allocation, r_u, $u \in U$, and a fixed way of scheduling the patient flows according to the policies described in Section 2.3.5 in Chapter 2.

The prediction horizon refers to the number of period in the future for which the prediction is made. The prediction horizon is denoted by h with $h \in \mathbb{N}_0$ and is expressed in units of days. At the beginning of day t_i we are interested in the predicted resource occupancy at the different hospital care units during the period $[t_i, t_{i+h}]$ depending on the admission scheme $\mathbf{a}_{[t_i,t_{i+h}]}$. This approach is also depicted in Figure 3.1. Since the resource usage at a unit fluctuates during a day, we model the daily resource usage as a probability distribution. Let $F^u_{t_j;\mathbf{a}_{[t_i,t_{i+h}]}}$ be the cumulative predicted resource-occupancy distribution function at unit u on day $t_j \in [t_i, t_{i+h}]$ determined at time t_i given admission scheme $\mathbf{a}_{[t_i,t_{i+h}]}$.

We analyze the predicted resource usage in terms of the occupancy probability distribution. In addition, we consider the quantiles of the resource-occupancy probability distribution as descriptive statistics to describe the main features of the distributions and use them to control patient admissions. The quantile values have been chosen as they provide key information on the underlying probability distribution. Moreover, quantiles are less susceptible to long-tailed distributions and outliers than for example the mean or other moment-related statistics [45]. In the following, the q-quantile of $F^u_{t_j;\mathbf{a}_{[t_i,t_{i+h}]}}$ is denoted by $F^{-1;u}_{t_j;\mathbf{a}_{[t_i,t_{i+h}]}}(q)$. The q-quantile value denotes the cut-off point where the number of resources used on day t_i is below $F^{-1;u}_{t_j;\mathbf{a}_{[t_i,t_{i+h}]}}(q)$ in q percent of the time. An example for a resource-occupancy distribution and corresponding quantile is presented in Figure 3.2. Here, $F^{-1;u}_{t_j;\mathbf{a}_{[t_i,t_{i+h}]}}(0.9)$ equals 13 which means that in 90% of the time on day t_i 13 or less beds are used and only in 10% of the time more than 13 beds are occupied (depicted by light-gray bars). Thus, using the q-quantile as descriptive statistic for decision support on admission control enables to control the risk of over-/under-utilization of resources caused by an admission scheme.

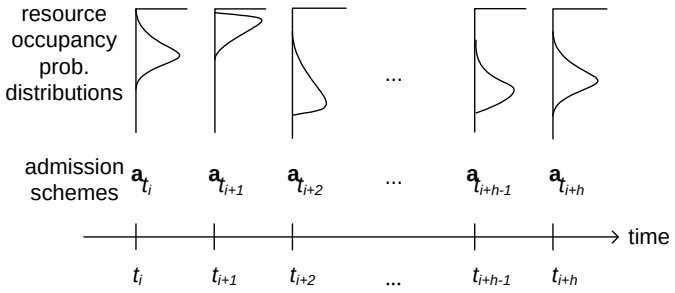


Figure 3.1: Example for the resource-occupancy distributions resulting from different admission decisions over time

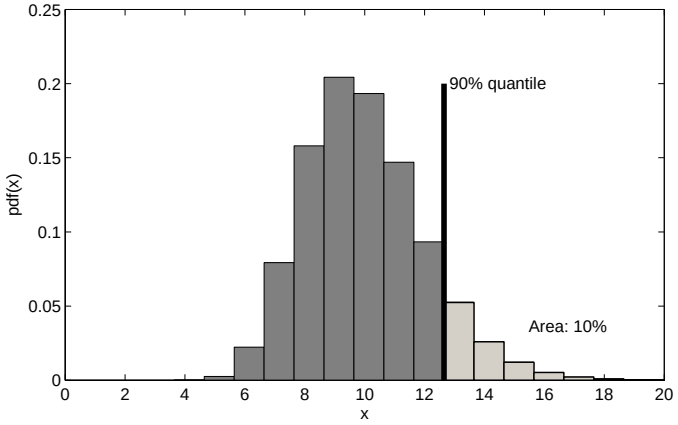


Figure 3.2: Histogram of an example bed usage distribution with 90%-quantile

3.4 Prediction by forward simulation

In the following section the method of forward simulation is described which is used for *what-if* simulations, i.e. simulations performed to examine the effect of an admission scheme on the future bed usage.

3.4.1 Approach

Forward simulation refers to the simulation of future time intervals given the current hospital admissions and an admission scheme for a sample of scenarios. Formally, forward simulation starting at time t_i consists of $N^{scenarios}$ simulation runs of the period $[t_i, t_{i+h}]$ given admission scheme $\mathbf{a}_{[t_i,t_{i+h}]}$. Forward simulation can thus be considered as Monte Carlo-sampling of the unknown underlying bed usage probability distribution. During a forward simulation run the following steps are performed:

1. Clone current system state at time t_i denoted by $\mathbf{s}_{t_i}$
2. Perform simulation steps given admission scheme $\mathbf{a}_{[t_i,t_{i+h}]}$ using $N^{scenarios}$ different random seeds
3. Reset system state $\mathbf{s}_{t_i}$

The system state, $\mathbf{s}_{t_i}$, is specified by the admission information of the patients that are already admitted to the different hospital units. The admission information of a patient comprises the patient's group and the time the patients have already spent at the unit they are currently admitted to.

3.4.2 Predicting the resource-usage probability distribution

Using forward simulation we obtain sample data on the resource usage resulting from an admission scheme, given the system state. From the obtained data we determine the empirical cumulative distribution function (ECDF) which is used as an estimator for the unknown underlying occupancy distribution, $F^u_{t_j;\mathbf{a}_{[t_i,t_{i+h}]}}$, $u \in U$. Let $\hat{F}^u_{t_j;\mathbf{a}_{[t_i,t_{i+h}]},N^{scenarios}}$ be the discrete ECDF of resource occupancy at unit u which is derived from the sampled bed usage data at u during day t_j obtained from forward simulation of $N^{scenarios}$ different scenarios. Since the resource occupancy may fluctuate considerably during the day, we monitor the hourly occupancy in the simulation. The hourly occupancy is then used to calculate the number of hours that x resources were occupied at unit u on day t_j in forward simulation scenario $n = 1, \ldots, N^{scenarios}$ given $\mathbf{a}_{[t_i,t_{i+h}]}$, denoted by $b^u_{x,t_j,n;\mathbf{a}_{[t_i,t_{i+h}]}}$.

Consider, for example, the resource occupancy at the CTS-HC in the CHE case study, cf. Section 2.3.5. If on Monday all beds are empty until the surgeries end and four type I patients arrive at about 18:00, this yields $b^{\text{CTS-HC}}_{0,t_i,1;\mathbf{a}_{[t_i,t_{i+h}]}} = 18$, $b^{\text{CTS-HC}}_{4,t_i,1;\mathbf{a}_{[t_i,t_{i+h}]}} = 6$ and $b^{\text{CTS-HC}}_{x,t_i,1;\mathbf{a}_{[t_i,t_{i+h}]}} = 0$ for $x = 1, 2, 3$.

$\hat{F}^u_{t_j;\mathbf{a}_{[t_i,t_{i+h}]},N^{scenarios}}$ can be calculated using the following non-parametric maximum likelihood estimator [53]:

$$\hat{F}^u_{t_j;\mathbf{a}_{[t_i,t_{i+h}]},N^{scenarios}}(y) = \frac{1}{N^{scenarios} \cdot 24} \sum_{n=1}^{N^{scenarios}} \sum_{x \le y} b^u_{x,t_j,n;\mathbf{a}_{[t_i,t_{i+h}]}}, \quad (3.1)$$

with $y \in \mathbb{N}_0$ and $t_j \in [t_i, t_{i+h}]$.

Thus, the ECDF is a consistent and unbiased estimator of the underlying unknown occupancy distribution [45].

The estimated q-quantile corresponding to $\hat{F}^{u}_{t_j;\mathbf{a}_{[t_i,t_{i+h}]},N^{scenarios}}$ is denoted by $\hat{F}^{-1;u}_{t_j;\mathbf{a}_{[t_i,t_{i+h}]},N^{scenarios}}(q)$. In the CTS-HC example, the resulting 90% quantile equals 4.

3.4.3 Experimental evaluation

In the following section, we analyze the accuracy and precision of the ECDF estimator, $\hat{F}^{u}_{\mathbf{a}_{t_j},N^{scenarios}}(y)$, depending on the number of forward simulation scenarios $N^{scenarios}$. Moreover, we evaluate the ECDF estimators in an explorative fashion and compare the resulting distributions and predictions to models used in the literature.

Our evaluation is based on occupancy data that was generated using the simulation described in Chapter 2. We used different prediction horizons and varied the underlying resource allocations and admission schemes as outlined below.

Setup agent-based simulation

The simulation instance used for the evaluation is based on the case study performed at the Catharina hospital Eindhoven (CHE), the Netherlands, reported in detail in Section 2.3.5. The instance specifies the patient pathway parameters, the involved care units $u \in U$ and the decision policies of the respective care unit agents concerning patient (re-)transfers. In the evaluation presented in this chapter, we explicitly distinguish between type III patients admitted to IC and MC, denoted by III(IC) and III(MC), respectively. As described in Chapter 2, type III(MC) patient flow is simulated in an abstract way such that the number of beds occupied by MC patients is sampled at the start of every day which directly determines the number of beds that are available for type I and II patients. Comparable to the situation at the CHE, the limited resource availability at the MC unit requires the rerouting of type I and II patient flows to higher care level units. As admission scheduling also controls the resource availability at the MC for the other patient flows, type III(MC) admissions are explicitly included in the admission scheme. Thus, Θ = {type I, type II, type III(IC), type III(MC), type IV}.

The occupancy data was generated using 50 independent simulation runs of 16 weeks including 4 weeks of warming-up and a prediction horizon, h, ranging between 0 and 7 days. Moreover, we varied the resource allocation in the simulation. We considered the current CHE resource allocation, $\mathbf{r}^{CHE}$,

given in Table 2.1, p. 45 and considered linear variations thereof that are denoted by $\mathbf{r}^{CHE\pm 20\%}$ with

$$\mathbf{r}^{CHE\pm 20\%} = (\lfloor r_u^{CHE} \cdot (1 \pm 0.2) + 0.5 \rfloor,\ u \in U). \tag{3.2}$$

Unconstrained admission control In this situation the admission schemes were determined based on the resource allocation employed in the simulation instance. The scheme was determined based on preliminary simulation runs as given in Table 3.1. The number of patient admissions of type I and II is limited by the available OR and postoperative care capacity, cf. Chapter 2. An upper limit of twice the number of allocated resources appeared more than sufficient for type III and IV patients, respectively.

$\mathbf{a}_{t_i}$	**Value**
$a_{t_i}^{I}$	$\min\{r_{\text{CTS-HC}}, r_{\text{CTS-OR}}\}$
$a_{t_i}^{II}$	$r_{\text{CTS-PACU}}$
$a_{t_i}^{III(IC)}$	$2 \cdot r_{\text{IC}}$
$a_{t_i}^{III(MC)}$	$2 \cdot r_{\text{MC}}$
$a_{t_i}^{IV}$	$2 \cdot r_{\text{IC-HC}}$

Table 3.1: Admission scheme for unconstrained admission control in the evaluation

Constrained admission control To assess the prediction performance in the case of constrained admission control, we employ an explorative method to generate many interesting states in the simulation. Specifically, we use a local search algorithm to randomly generate admission schemes for which we predict the resulting occupancy distributions and to choose one of the candidate admission scheme to be employed in the subsequent period in the simulation. This allows for a large variety of occupancy distributions that we use for our analysis of our prediction approach.

The admission schemes employed in a simulation run during the time period $[t_i, t_{i+h}]$, $i = 1, 2, \ldots$ are iteratively determined using a hill-climbing search algorithm with the objective to achieve a maximal patient throughput, given a balanced patient mix, being a popular method for hard, practical problems [82] which is the case here. Candidate admission schemes are generated by randomly adjusting the currently employed admission scheme. The occupancy data was collected given the currently employed admission scheme as well as the evaluated candidate admission schemes.

The hill-climbing search algorithm that we used in our experiments is described in pseudo-code notation in Algorithm 1. The initial admission

scheme at time t_0 is determined based on the employed resource allocation, r_u, $u \in U$, and the patient arrival processes as given in Table 3.2. Similarly to the unconstrained admission control case, the number of patient admissions of type I and II is initially determined by the available OR and postoperative care capacity. For type III and IV patients the initial number of admissions is set to 2 and 3, respectively, which corresponds to the mean arrival rates, cf. Section 2.3.5.

At time t_i, the currently employed admission scheme, $\mathbf{a}_{[t_{i-1},t_{i+h-1}]}$, is used as default solution based on which the method `generateNeighbors` generates candidate schemes by randomly in- or decreasing the number of patient admissions in the default admission scheme. The in- or decrement is sampled randomly from a uniform distribution such that the number of patients to be admitted does not exceed the number of resources allocated to the unit to accommodate the patients. In order to refine the search, the bounds for in- and decreasing the number of patient admissions and the number of generated candidate solutions are decreased exponentially in the course of the hill-climber run. The hill-climbing algorithm is terminated after a predefined number of iterations.

$\mathbf{a}_{t_0}$	**Value**
$a^{I}_{t_0}$	$\min\{r_{\text{CTS-HC}}, r_{\text{CTS-OR}}\}$
$a^{II}_{t_0}$	$r_{\text{CTS-PACU}}$
$a^{III(IC)}_{t_0}$	2
$a^{III(MC)}_{t_0}$	3
$a^{IV}_{t_0}$	3

Table 3.2: Initial admission scheme for constrained admission control

The method `evaluateScheme` evaluates an admission scheme using forward simulation. The method returns a vector of predicted q-quantiles for the bed usage at the different care units. A candidate scheme $\mathbf{a}'_{[t_i,t_{i+h}]}$ has to satisfy the resource constraint given by

$$\hat{F}^{-1;u}_{t_j;\mathbf{a}'_{[t_i,t_{i+h}]}}(q) \leq r_u, \forall u \in U, j = i, \ldots, i+h. \tag{3.3}$$

In order to achieve a balanced patient mix the hill-climbing algorithm maximizes the minimal number of patient admissions of the scheme, $MPA(\mathbf{a}_{[t_i,t_{i+h}]})$, is evaluated which is given by

$$MPA(\mathbf{a}_{[t_i,t_{i+h}]}) = \min_{g \in \Theta} \sum_{k=i}^{i+h} a^{g}_{t_k}. \tag{3.4}$$

Initial experiments showed that using the total number of patient admissions as objective function to be maximized resulted in a biased selection of type II admissions where the corresponding pathways feature few postoperative complications and comparatively short LoS. In general, such admission schemes are undesirable and unrealistic for hospitals where this would lead to excessive waiting lists for more "complex" patients. Therefore, $MPA(\mathbf{a}_{[t_i,t_{i+h}]})$ was chosen as objective function which provides for a more balanced patient mix in the simulation. If no candidate solution satisfies the resource constraint (3.3), the hill-climbing algorithm uses the default admission scheme as a solution. Since the optimization is performed online during simulation, a sub-optimal default solution chosen now solely affects the current time step and is likely to be improved in following step of the optimization. The parameter settings were determined in preliminary simulation experiments and were set to 4 iterations of the hill-climbing algorithm. The adjustment parameters and the number of solutions to be generated is are decreased exponentially with a rate set to 0.25. During preliminary experiments these values appeared to allow for a sufficiently extensive local search in a reasonable runtime.

The data presented in the following sections contains the randomly generated candidate and employed admission schemes as well as the resulting resource-occupancy distribution.

Accuracy and precision of the ECDF estimator

In general, the accuracy of a prediction is the degree of closeness of the estimates to the true value. Using the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality [31], we can determine the number forward simulation scenarios needed to obtain a prediction of the underlying bed occupancy distribution with predefined accuracy. Specifically, the DKW inequality bounds the probability that $\hat{F}^u_{t_j;\mathbf{a}_{t_j,N^{scenarios}}}(y)$ differs from $F^u_{t_j;\mathbf{a}_{[t_i,t_{i+h}]}}(y)$ by more than a given constant $\epsilon > 0$ for any $y \in \mathbb{N}$. Formally, the DKW inequality states that

$$Pr(\sup_y |\hat{F}^u_{\mathbf{a}_{t_j,N^{scenarios}}}(y) - F^u_{t_j;\mathbf{a}_{[t_i,t_{i+h}]}}(y)| > \epsilon) \leq 2e^{-2N^{scenarios}\epsilon^2} \; \forall \epsilon > 0, \tag{3.5}$$

where $e^{-2n\epsilon^2} \leq 0.5$ If we want to ensure that the possible error of $\hat{F}^u_{\mathbf{a}_{t_j,N^{scenarios}}}$ is at most $\epsilon = \frac{1}{10}$, with at least 90% confidence, we get for $N^{scenarios}$ that

$$2e^{-2N^{scenarios}/100} \leq 1 - 0.9 \Leftrightarrow N^{scenarios} \geq 50 \ln 20 \approx 150.$$

Algorithm 1: Pseudo-code description of hill-climbing local search algorithm for constrained admission control

Input: Scheme $\mathbf{a}_{[t_{i-1},t_{i+h-1}]}$
Result: Scheme $\mathbf{a}_{[t_i,t_{i+h}]}$

`//initialization`
for $j = i$ **to** $i + h - 1$ **do**
    $\mathbf{a}_{t_j}^{default} = a_{t_j}$;
$\mathbf{a}_{t_{i+h}}^{default} = a_{t_{i+h-1}}$;
`//optimization`
$nextEval = -\infty$;
$nextSolution = \mathbf{a}_{[t_i,t_{i+h}]}^{default}$;
repeat
    $\mathcal{A} =$ `generateNeighbors(`*nextSolution, currentIteration*`)`;
    forall $\mathbf{a}' \in \mathcal{A}$ **do**
        $\hat{F}^{-1}_{t_j;\mathbf{a}'}(q) =$ `evaluateScheme` $(\mathbf{a}')$;
        if $\hat{F}^{-1;u}_{t_j;\mathbf{a}'}(q) \leq r_u \,\forall$ *units* $u \in U$, $j = i, \ldots, i + h$ **then**
            if $MPA(\mathbf{a}') >$ *nextEval* **then**
                $nextSolution = \mathbf{a}'$;
                $nextEval = MPA(\mathbf{a}')$;
until *termination* ;
return nextSolution;

Note, that using $\hat{F}^{u}_{\mathbf{a}_{t_j,N^{scenarios}}}$ we do not need to make any assumptions on the unknown underlying distribution $F^{u}_{t_j;\mathbf{a}_{[t_i,t_{i+h}]}}$.

The precision of a prediction relates to the exactness of the operation used to obtain the estimates. To assess the precision of the ECDF for different number of forward simulation scenarios we determine the empirical confidence interval of the maximum likelihood estimator using Greenwood's formula [36]. In Figure 3.3 an example empirical cumulative distribution function is shown with upper and lower limits or bounds of the 95% confidence interval. The precision of the estimator is defined in terms of the maximum norm of the length of the estimates' 95% confidence intervals. The maximum norm assigns to a real-valued bounded function f the nonnegative number

$$||f||_\infty = \max\{|f(x)| : x \in \text{domain of } f\}.$$

We define the maximum norm of the length of the 95% confidence interval

of $\hat{F}^u_{\mathbf{a}_{t_j,N^{scenarios}}}$ as the maximum norm of the difference between the 97.5% and 2.5% confidence bounds, i.e.

$$||\text{ci}||_\infty = \max_{x \geq 0}\{|2z_{0.025}\sqrt{\widehat{Var}[\hat{F}^u_{\mathbf{a}_{t_j,N^{scenarios}}}(x)]}|\}, \tag{3.6}$$

where z_q denotes the q-quantile of the standard normal distribution with $z_{0.025} = -z_{0.975} = -1.96$. $\widehat{Var}[\hat{F}^u_{\mathbf{a}_{t_j,N^{scenarios}}}(x)]$ denotes the sample variance of $\hat{F}^u_{\mathbf{a}_{t_j,N^{scenarios}}}(x)$ according to [36].

Using this precision metric, we get a worse-case bound on the overall precision of the calculated estimator. Thus, a small $||\text{ci}||_\infty$ value, i.e. a small confidence interval, yields a high precision of the bed usage distribution estimator. In order to determine the number of forward simulation scenarios needed to obtain accurate prediction results we calculated the mean $||\text{ci}||_\infty$ value of the $\hat{F}^u_{\mathbf{a}_{t_j,N^{scenarios}}}$ estimates for the IC, MC, HC-IC and CTS-HC units and the CTS-ward for varying $N^{scenarios}$ values.

Unconstrained patient admission scheduling The results for the precision of the ECDF estimator for the IC are shown in Figure 3.4. The number of simulation runs, $N^{scenarios}$, varies between 10 and 10^3 and the prediction horizon is set to $h = 0, \ldots, 7$. From Figure 3.4 we can conclude that the confidence bounds of the ECDF converge polynomially for increasing $N^{scenarios}$. Moreover, the figure shows the small impact of the length of the prediction horizon h on the precision of the estimator. For $h = 0$ the estimator shows slightly less variability compared to prediction horizons with $h \geq 1$ days. This result can be attributed to the randomness of patient admissions in this situation. Patient arrivals are solely limited by the patients' actual demand for the types of care provided by the units and the available resource capacity which fluctuates over time. The resource availability for $h = 0$, however, shows less variability since the resource occupancy is partly determined by prior patient admissions and transfers, with an average LoS of more than 1 day, which is why less variability is present in the $h = 0$ data compared to the $h \geq 1$ data. In a sensitivity analysis we analyzed the effect of the resource allocation at the hospital units on the estimates' $||\text{ci}||_\infty$ values which showed that an in-/decrease in resource capacity of 20% resulted in a maximal relative difference of less than 10^{-3} for the IC. Thus, the convergence appears to be robust for different resource allocations. The convergence and sensitivity results are comparable for the other units which are therefore omitted here.

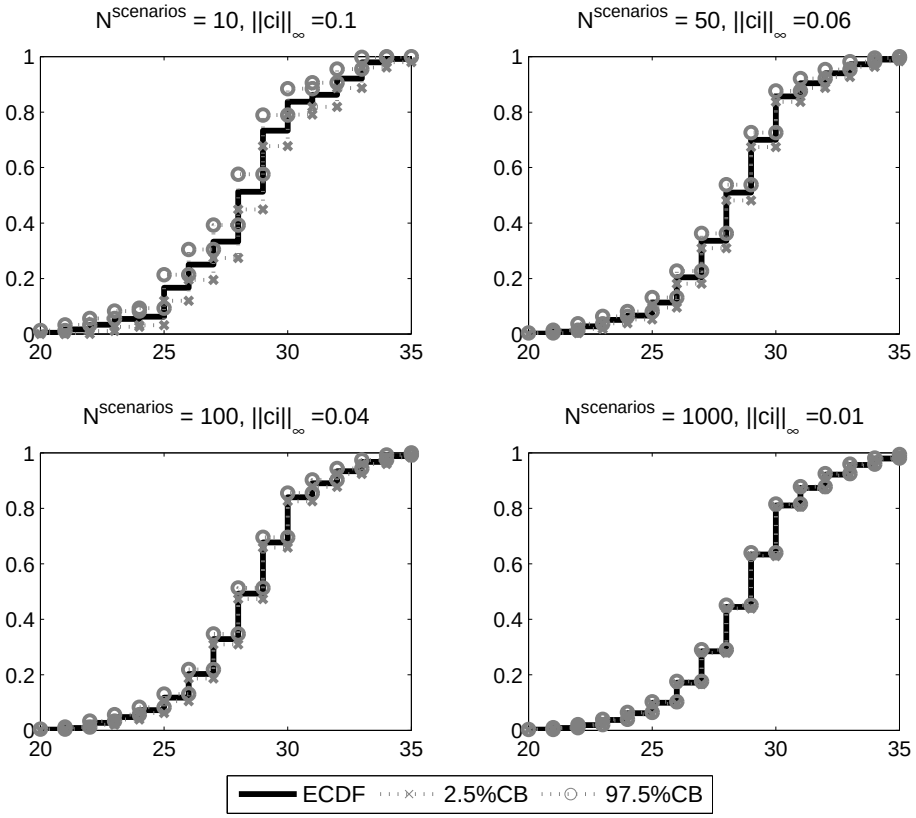


Figure 3.3: ECDF with lower (2.5%) and upper (97.5%) confidence bounds (CB) determined by (3.6) for $N^{scenarios} = 10, 50, 100, 1000$ with corresponding $||\text{ci}||_\infty$ values

Constrained patient admission control The precision of the ECDF estimator for constrained admission control is shown in Figure 3.5 for varying $N^{scenarios}$ and a prediction horizon ranging between $h = 0$ and 7 days on a log-log scale. The convergence of the ECDF is comparable to the unconstrained admission control case with polynomially converging confidence bounds for increasing $N^{scenarios}$. In contrast to the unconstrained case, however, an increasing prediction horizon h in constrained admission control causes larger confidence intervals. For the CTS-ward data, the increase in variability is marginal, while for the MC and IC a larger increase is to be noted which remains almost constant for $h \geq 1$. For the CTS-HC and IC-HC data, a prediction horizon of $h > 0$ also causes larger confidence intervals, however, no consistent in- or decrease in variability is to be seen which may be attributed to the small size of the care units, the complex patient routing and arrivals of type I and IV patients as well as the variability

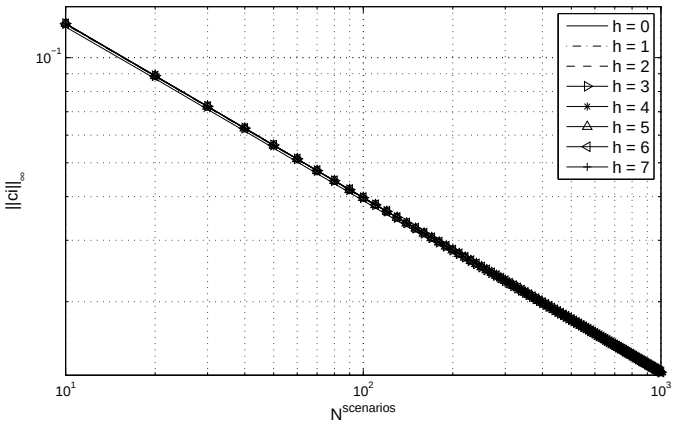


Figure 3.4: Average $||\text{ci}||_\infty$ value for varying $N^{scenarios}$ and $h = 0, \ldots, 7$ on a log-log scale for the IC

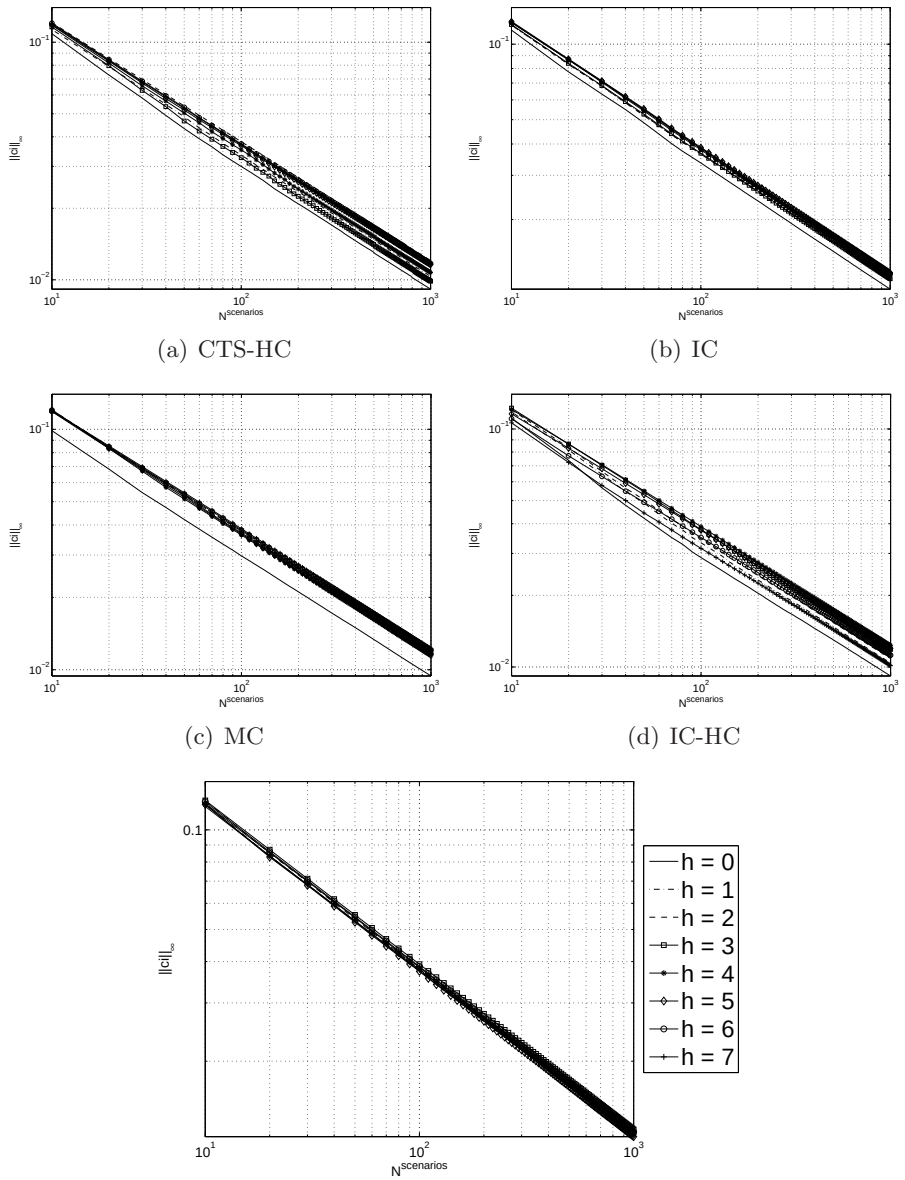
of the corresponding patient pathways, cf. Section 2.3.5.

Determining the number of forward simulation runs

On an Intel Pentium 4 2.4GHz machine with 2GB RAM a single forward simulation run takes about 13.944 and 4.34 seconds on average for $h = 7$ days for $N^{scenarios} = 1000$ and $N^{scenarios} = 300$, respectively. At the same time, the overhead of computational time needed for copying the current state information at the different units, resampling patient paths, etc. remains constant. For $N^{scenarios} = 300$, the $||\text{ci}||_\infty$ values of the different units for the (un-)constrained admission control case are about 0.02 which means that the estimates have minimal variability. For $N^{scenarios} = 1000$, the $||\text{ci}||_\infty||$ values of the different units for the (un-)constrained admission control case are about 0.01. In consideration of the decrease of less about $8 \cdot 10^{-3}$ of the already small $||\text{ci}||_\infty$ values, the computational effort of the additional simulation runs greatly outweighs the gain in precision. Therefore, the number of forward simulation runs is set to $N^{scenarios} = 300$ on which the following results are based.

Analysis of the empirical distribution functions of hospital bed occupancy

Next, we contemplate the shape of the resource-usage probability distributions obtained from forward simulation for unconstrained and constrained patient admission control. In addition to the basic resource allocation, we



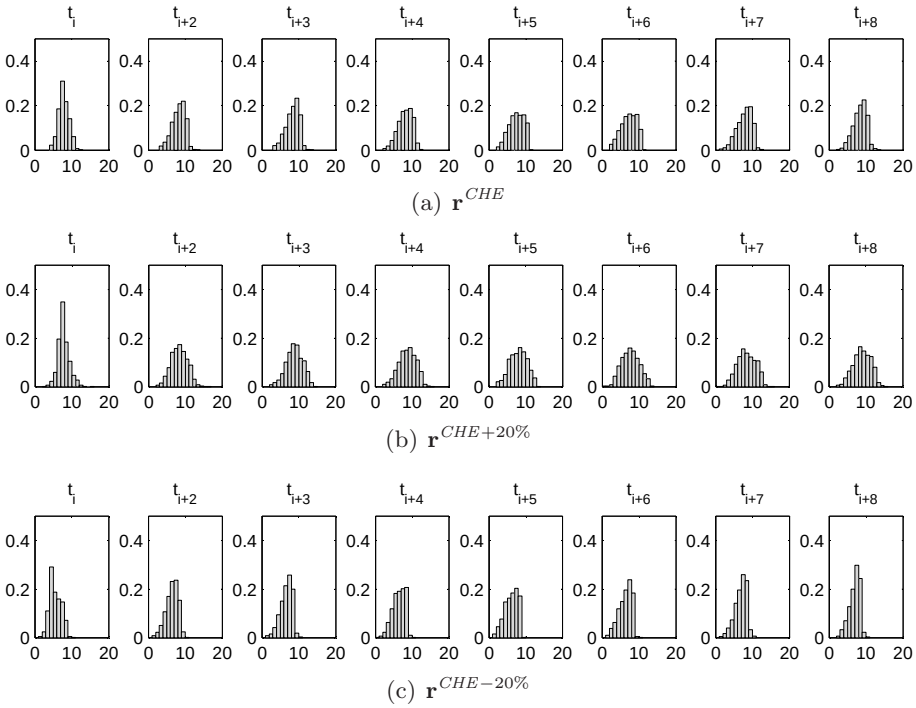
(a) CTS-HC (b) IC

(c) MC (d) IC-HC

(e) CTS-ward

Figure 3.5: Average $||\text{ci}||_\infty$ values (y-axes) for $N^{scenarios} = 10, 20, \ldots, 1000$ (x-axes) and $h = 0, \ldots, 7$ on a log-log scale for (a) CTS-HC, (b) IC, (c) MC, (d) HC-IC and (e) CTS-ward; the legend in (e) also applies to (a) - (d)

also performed a sensitivity analysis to study the robustness of the obtained distributions.



(a) $\mathbf{r}^{CHE}$

(b) $\mathbf{r}^{CHE+20\%}$

(c) $\mathbf{r}^{CHE-20\%}$

Figure 3.6: Histograms of resource-occupancy ECDF for unconstrained admission control for (a) basic allocation, (b) basic allocation +20% and (c) basic allocation -20% at IC

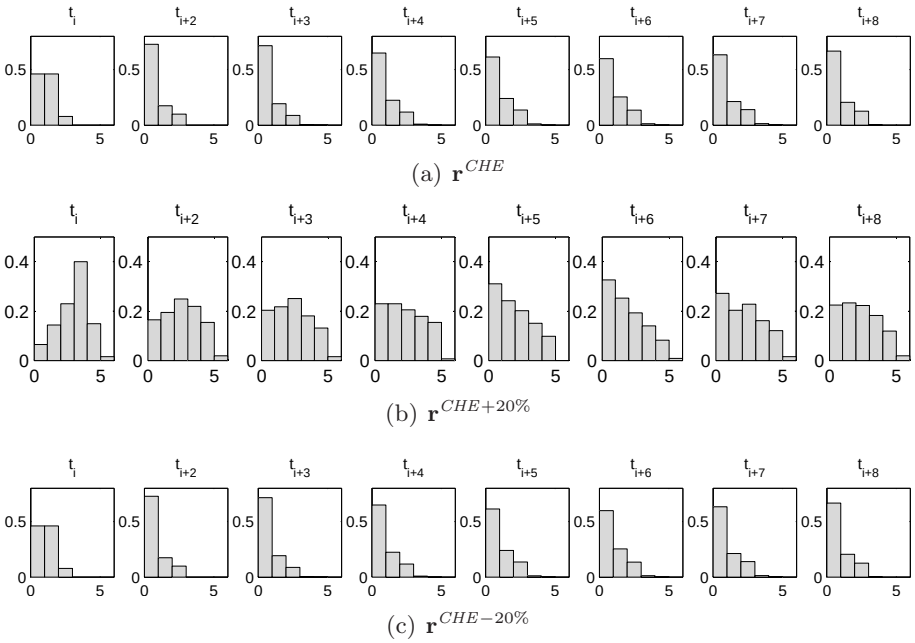
Unconstrained patient admission control Figure 3.6, 3.7, 3.8, 3.9 and 3.10 show histograms of the resource-usage ECDF at IC, MC, IC-HC, CTS-HC and CTS-ward during the period $t_i, \ldots, t_{i+7}$ days for the considered resource allocations.

For all allocations shown in Figure 3.6 the resource-usage distribution tends to be right-skewed with the mass of the distribution being concentrated on the left of the figures. Analogous observations can be made for Figure 3.7, 3.8 and 3.10. Especially for the small units, the MC, IC-HC and CTS-HC, the discreteness of the distribution support and values due to the arrival schemes provides for a large range of distribution shapes, e.g. binomial, Poisson and almost uniform distributions where all possible values of resource occupancy are almost equally likely. For the CTS-ward the distribution shapes are more or less symmetrical.

To assess the dispersion of the measured ECDFs, we consider the 90% quantile ranges, defined as the difference between the 5% and 95% quantiles. Table 3.3 contains the 90% quantile ranges averaged over the data samples.

Unit	allocation	$h=0$	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$
CTS-HC	$\mathbf{r}^{CHE}$	3.1964	3.4286	3.4286	3.4286	3.4286	3.4286	3.4286	3.4286
	$\mathbf{r}^{CHE+20\%}$	4.1071	4.2857	4.2857	4.2857	4.2857	4.2857	4.2857	4.2857
	$\mathbf{r}^{CHE-20\%}$	2.3571	2.5357	2.5714	2.5714	2.5714	2.5714	2.5714	2.5714
IC	$\mathbf{r}^{CHE}$	3.5625	4.2857	4.3036	4.4107	4.415	4.4286	4.4321	4.4464
	$\mathbf{r}^{CHE+20\%}$	4.6607	6.3036	6.8036	6.9464	7.0536	7.125	7.1786	7.3214
	$\mathbf{r}^{CHE-20\%}$	4.0714	5.25	5.5357	5.6071	5.6357	5.6786	5.6964	5.6964
MC	$\mathbf{r}^{CHE}$	3.7321	3.9643	3.9821	3.9821	4.0	4.0	4.0	4.0
	$\mathbf{r}^{CHE+20\%}$	4.1786	4.6964	4.8393	4.8036	4.8571	4.875	4.893	4.905
	$\mathbf{r}^{CHE-20\%}$	2.9107	3.0	3.0	3.0	3.0	3.0	3.0	3.0
IC-HC	$\mathbf{r}^{CHE}$	2.5893	3.1429	3.1071	3.2143	3.2321	3.25	3.2589	3.2679
	$\mathbf{r}^{CHE+20\%}$	2.9464	3.8214	3.8929	3.8929	3.9464	3.9821	3.9821	3.9875
	$\mathbf{r}^{CHE-20\%}$	1.9821	2.6429	2.6786	2.6875	2.7143	2.7286	2.7407	2.764
CTS-ward	$\mathbf{r}^{CHE}$	8.0536	9.1429	9.5179	9.6875	9.9821	10.5179	10.9107	11.1071
	$\mathbf{r}^{CHE+20\%}$	9.9107	11.0357	11.3482	11.7232	11.8875	12.1161	12.3839	12.4107
	$\mathbf{r}^{CHE-20\%}$	6.0625	6.8929	7.25	7.4196	7.9821	8.5536	9.0536	9.2946

Table 3.3: Average 90% quantile range of ECDF at different units for different resource allocations



(a) $\mathbf{r}^{CHE}$

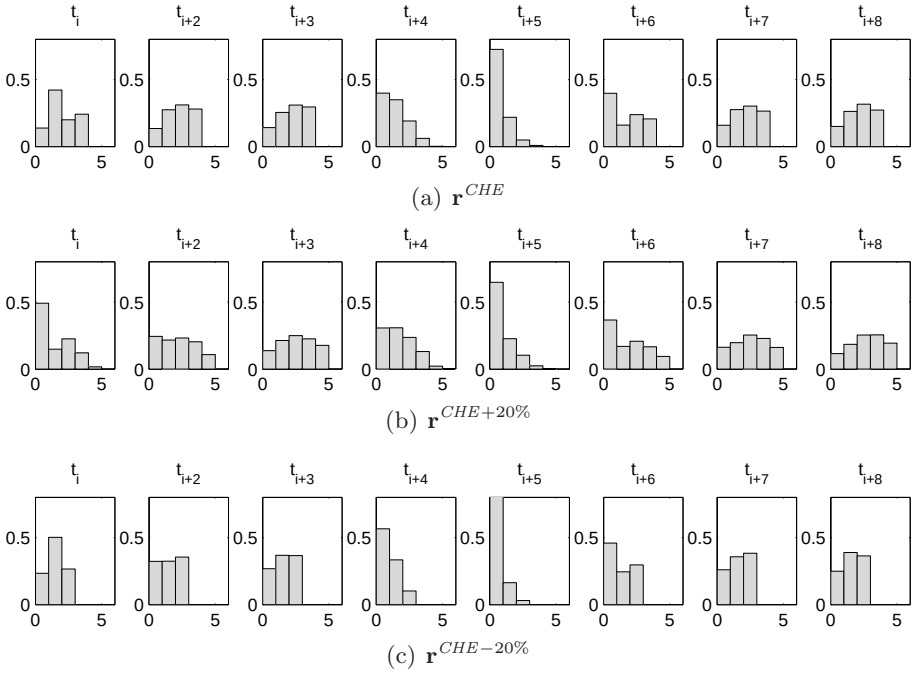
(b) $\mathbf{r}^{CHE+20\%}$

(c) $\mathbf{r}^{CHE-20\%}$

Figure 3.7: Histograms of resource-occupancy ECDF for unconstrained admission control for (a) basic allocation, (b) basic allocation +20% and (c) basic allocation -20% at MC

The average quantile ranges show that for a given resource allocation the dispersion of the ECDF increases for increasing h. For the large units, i.e. the IC and CTS-ward, the ECDF show very small dispersion relative to the number of allocated resources whereas the ECDFs of small units show a larger dispersion that can be attributed to the discreteness of the possible values and the large variability of LoS at the units, cf. Section 2.4.1. Except for the CTS-ward, the dispersion remains almost constant for $h > 0$.The overall increase in dispersion of the ECDF is to be expected since prediction precision generally decreases with increasing time horizon, especially for complex and dynamic forecasting problems as with resource-usage prediction. This is also the reason for the larger values of $||\text{ci}||_\infty$ for increased prediction horizon h discussed earlier in this section.

In Table 3.4 the variability of the predicted q-quantile values is shown for a prediction horizon h ranging between 1 and 7 days. The variability of the predicted forecasts is defined as the mean absolute difference between the q-quantile predictions for t_k determined at t_{k-h}. From Table 3.4 we can conclude that the variability of quantile predictions is small for the different units and quantile values. The largest difference in predicted q-quantile



(a) $\mathbf{r}^{CHE}$

(b) $\mathbf{r}^{CHE+20\%}$

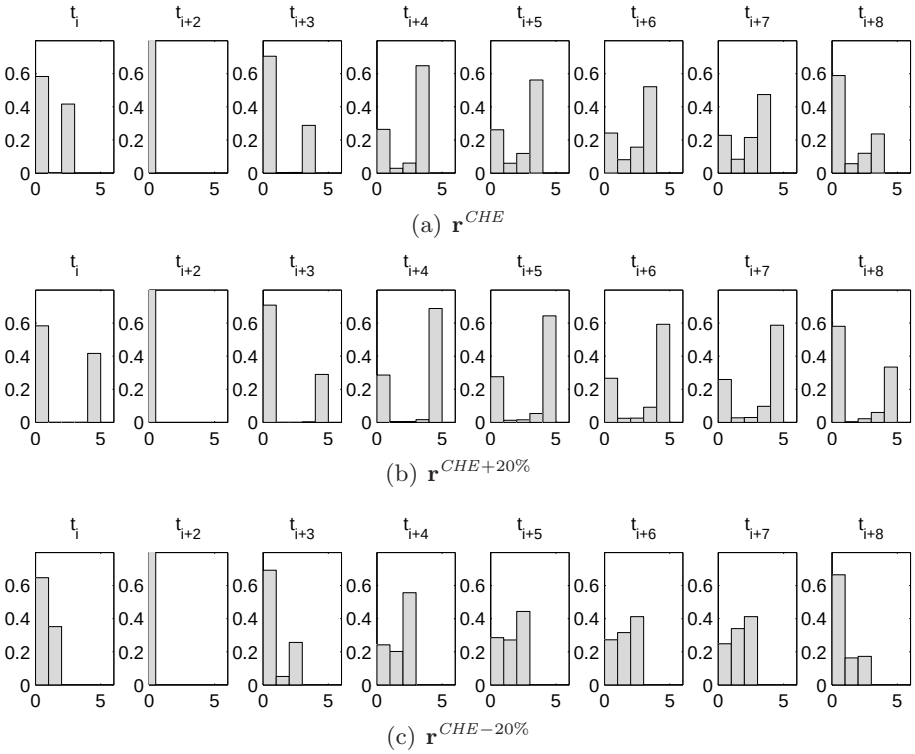
(c) $\mathbf{r}^{CHE-20\%}$

Figure 3.8: Histograms of resource-occupancy ECDF for unconstrained admission control for (a) basic allocation, (b) basic allocation +20% and (c) basic allocation -20% at IC-HC

Unit	q	h = 1	h = 2	h = 3	h = 4	h = 5	h = 6	h = 7
IC	70%	0.9184	1.0	1.0	1.0306	1.0816	1.102	1.102
	80%	0.7551	0.8571	0.8571	0.9187	0.9187	0.9796	0.9796
	90%	0.4898	0.6122	0.6327	0.6735	0.6827	0.6939	0.6939
IC-HC	70%	0.2041	0.2137	0.2245	0.2245	0.3061	0.2857	0.2653
	80%	0.3061	0.3265	0.3265	0.3469	0.3469	0.3673	0.4082
	90%	0.2653	0.2653	0.2653	0.2653	0.2653	0.2653	0.2653
MC	70%	0.6122	0.6122	0.5714	0.5918	0.5918	0.5918	0.5918
	80%	0.5714	0.602	0.6122	0.6327	0.6327	0.6327	0.6327
	90%	0.102	0.102	0.102	0.102	0.102	0.102	0.102
CTS-ward	70%	1.0306	1.1959	1.2939	1.4184	1.6837	1.7143	1.8571
	80%	1.1837	1.4898	1.499	1.6122	1.755	1.7747	1.8163
	90%	1.1837	1.5122	1.6102	1.7143	1.8796	1.8980	1.9571

Table 3.4: Variability of predicted q-quantiles at different units over time

over time can be observed for the CTS-ward where the predictions differ on average by less than 2 beds. For the IC, IC-HC and MC the greatest difference in predictions is limited to about 1, 0.4 and 0.63 beds, respectively.



(a) $\mathbf{r}^{CHE}$

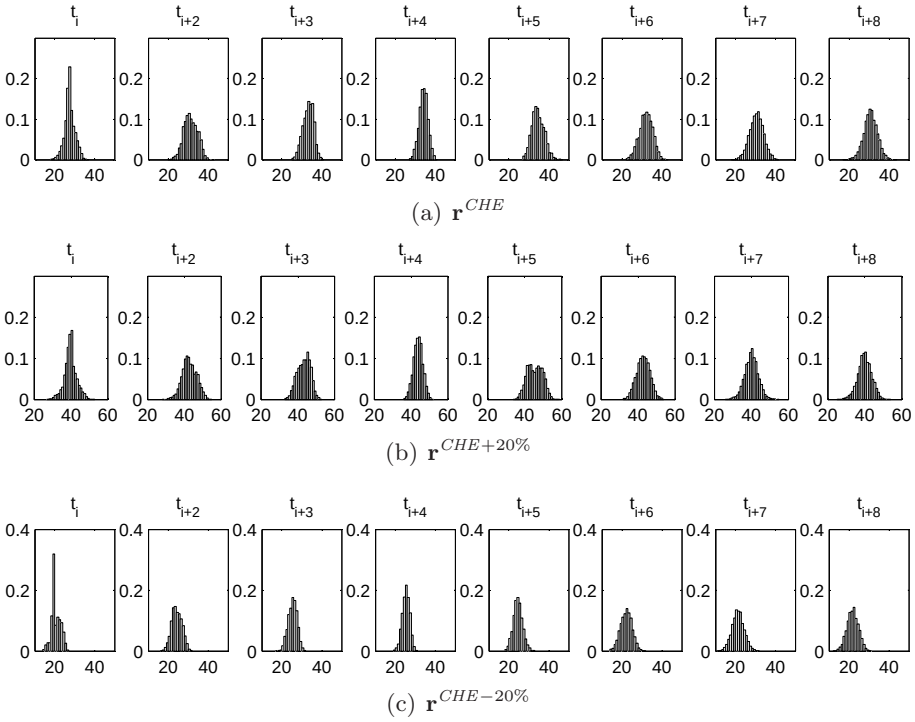
(b) $\mathbf{r}^{CHE+20\%}$

(c) $\mathbf{r}^{CHE-20\%}$

Figure 3.9: Histograms of resource-occupancy ECDF for unconstrained admission control for (a) basic allocation, (b) basic allocation +20% and (c) basic allocation -20% at CTS-HC

The variability relative to the number of allocated resources decreases for increasing value of q. For the IC occupancy data, the maximal relative variability amounts to 12.2%, 8.9% and 5.3% for q equal to 0.9, 0.8 and 0.7, respectively. For the IC-HC, the variability decreases from maximally 10.2% to 5.3% and for the CTS-ward the decrease amounts to 6.6%, 5.1% and 4.7% for q equal to 0.9, 0.8 and 0.7, respectively. The maximal variability for the MC quantile predictions decreases from 20% to about 2% relative to the number of allocated resources. Here, the largest variability occurs which can be explained by the random patient arrivals in our simulation, cf. Section 2.4.1 and the small number of the possible occupancy values.

Constrained patient admission control Figure 3.11, 3.12, 3.13, 3.14 and 3.15 show histograms of the density functions for the ECDF of bed usage at the IC, MC, IC-HC, CTS-HC units and the CTS-ward resulting from 300 forward simulation scenarios with a prediction horizon of $h = 0, \ldots, 7$ days.



(a) $\mathbf{r}^{CHE}$

(b) $\mathbf{r}^{CHE+20\%}$

(c) $\mathbf{r}^{CHE-20\%}$

Figure 3.10: Histograms of resource-occupancy ECDF for unconstrained admission control for (a) basic allocation, (b) basic allocation + 20% and (c) basic allocation -20% at CTS-ward

The rows in the figures correspond to different system states at time t_i and the columns correspond to the different days in the interval $[t_i, t_{i+7}]$ with the admission decisions indicated at the top. The figures show that empirical bed usage distributions can differ considerably depending on the state and the admission decisions over time. Overall, the distribution shapes vary from skewed to almost symmetrical distributions with varying peakedness. For the CTS-ward even bimodal distributions can be observed which can be explained by canceled surgeries and thus prolonged admission as a result of the blocking of postoperative care units. Therefore, a summary statistic such as the mean, median and standard deviation may not be representative for the different admission schemes and decision moments.

From the depicted distributions it becomes apparent that the predictions differ considerably over time and depend on the employed admission schemes. As the data includes a large variety of admission schemes and their predicted resource occupancy, including admission schemes that do not maximize the objective (3.4) or fulfill the resource constraint (3.3), the

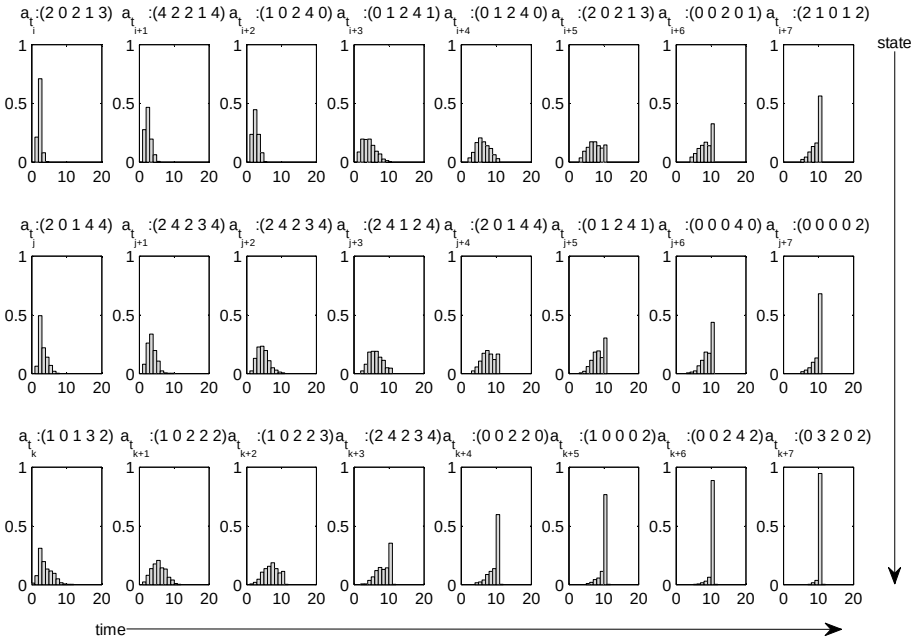


Figure 3.11: Histograms of resource-occupancy ECDF at IC resulting from different states and admission schemes

variability of the obtained resource-occupancy quantiles and thus the dispersion of the distributions is large. Moreover, as the admission schemes are adjusted continuously during simulation to maximize (3.4) and satisfy (3.3) according to Algorithm 1, a comparison of the predicted quantiles over time is not feasible.

The prediction horizon to be used will typically depend on the lead time of an admission decision, i.e. a long decision lead time would typically induce a long prediction horizon in order to account for resource shortage resulting from the decision. If the decision lead time is short, however, a shorter prediction horizon may be advisable due to the inherent variability of resource occupancy over time.

Comparison to utilization distributions in the literature

In the literature a Normal distribution is sometimes assumed for the resource usage due to the independent and identically distributed discrete random effects at a unit, cf. Kusters and Groot [60]. Our empirical simulation results show that the bed usage distributions are often non-symmetrical, especially for small units like the MC and in the case of constrained admission control.

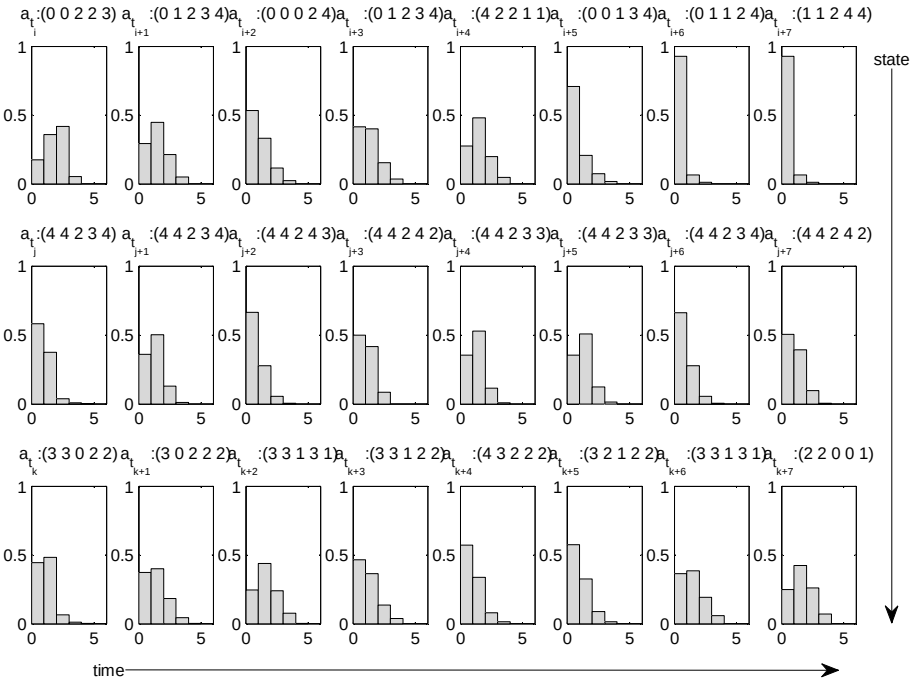


Figure 3.12: Histograms of resource-occupancy ECDF at MC unit resulting from different states and admission schemes

Moreover, applying the Shapiro-Wilk test of normality on the data obtained from the simulations the nullhypothesis of normality is rejected in more than 85% of the samples at a significance level of 5%. Using normal distributions for sampling LoS and arrival data would probably promote the normality of the resource-usage data. However, normally distributed LoS and arrival times are not realistic in this problem setting as LoS times are typically right-skewed[1] and arrival times are typically Poisson [63, 96]. Thus, in general the normality assumption will not hold.

However, if the mean is used as an estimate for resource occupancy resulting from an admission scheme or for capacity planning purposes, the normality assumption may be reasonable. Especially in the case of unconstrained admission control and larger care units, the distributions depicted above appear symmetrical. The mean of a normal distribution then provides a meaningful estimate to be used for optimizing the admission or capacity planning decisions for the bulk of cases. On the other hand, if quantile values are of interest for the optimization, the normality assumption may lead

[1]Note that in our simulation we employ Lognormal distributions.

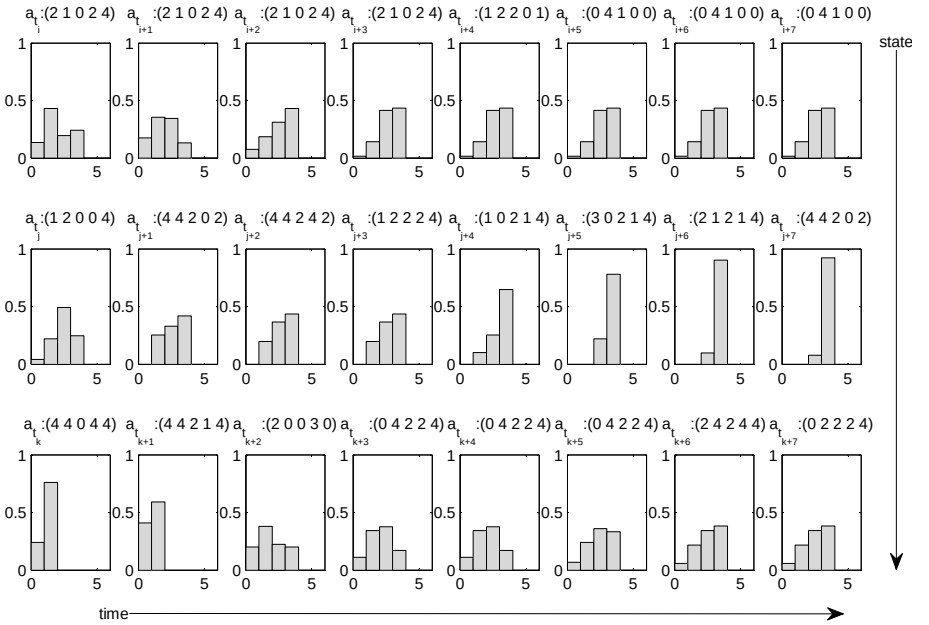


Figure 3.13: Histograms of resource-occupancy ECDF at IC-HC resulting from different states and admission schemes

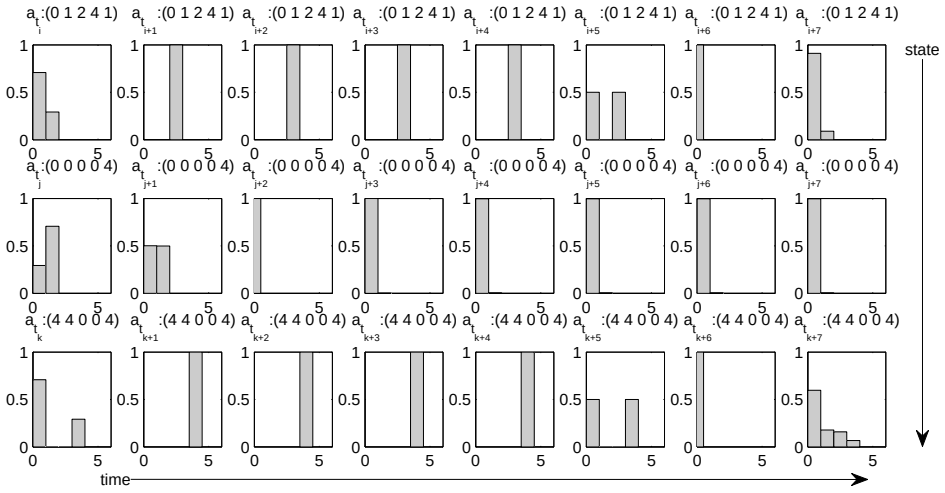


Figure 3.14: Histograms of resource-occupancy ECDF at CTS-HC resulting from different states and admission schemes

to significantly different estimates due to the tails of the normal distribution that are not present in the samples shown in the figures above.

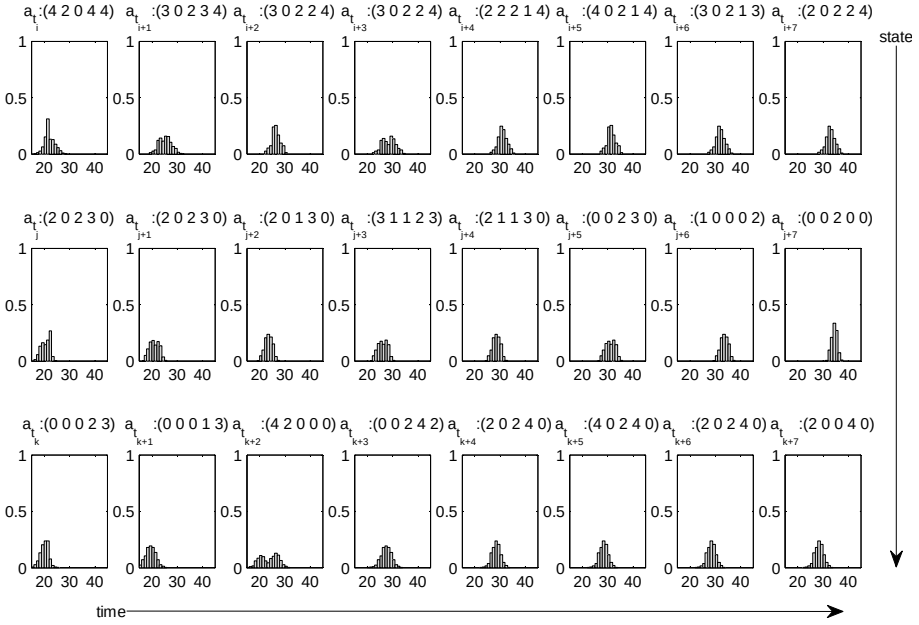


Figure 3.15: Histograms of resource-occupancy ECDF at CTS-ward resulting from different states and admission schemes

3.5 Prediction by supervised learning

In the face of the increased computational effort associated with prediction by forward simulation, we evaluated the possibility of using supervised learning techniques to approximate the resource-usage probability distribution. In general, in supervised learning we are given a set of example pairs (x_j, y_j), $x_j \in X$, $y_j \in Y, j = 1, 2, \ldots$ and the aim is to find a function $f : X \to Y$ in the allowed class of functions that best matches the examples. In other words, we wish to infer the mapping that underlies the data.

In this section, the resource-usage probability distribution at a care unit is to be predicted based on (aggregated) available system state information. In case of unconstrained admission control the units' resource occupancy is used as input. In case of constrained admission control the admission scheme is used in addition to the resource occupancy. The function is to be learned from the data obtained from forward simulation discussed in Section 3.4. The learned models are benchmarked against a basic heuristic derived from hospital practice. As concluded in Section 3.4, the variability and predictability of the resource occupancy decreases with increasing prediction horizon h. Therefore, we initially evaluated the supervised learning approach for $h = 1$.

We want to remark that the previous approach will be used in the remainder of the thesis, so readers can skip this section at first reading.

3.5.1 Approach

Similarly to the forward simulation approach presented in Section 3.4, our prediction approach by supervised learning focusses on predicting the resource-occupancy probability distribution. Thus, our aim is to learn a function for the resource-usage ECDF $F^u_{t_j:\mathbf{a}_{[t_i,t_{i+h}]}}$ for each unit $u \in U$. Predicting solely the value of a descriptive statistic like the mean or a quantile value would significantly decrease the complexity of the prediction problem and thus increase the performance of the supervised learner. However, due to the broad variety of encountered probability distributions in Section 3.4.3, selecting an appropriate descriptive statistics to be used for decision support is unclear as the choice would typically depend on the focus of decision support, the encountered ECDF shapes, etc. Predicting the resource-occupancy ECDF is a more generally applicable approach as it allows to derive mean occupancy values, quantile values or any other descriptive statistics of interest depending on the scope of decision support.

Based on preliminary results approximating $\hat{F}^u_{\mathbf{a}_{t_j,N^{scenarios}}}$ using an appropriate class of distribution functions and learn the parameters of this class using an artificial neural network (ANN) was found to be superior to other approaches like approximation using high-degree polynomials. Our approach avoids encountered oscillatory behavior especially for the distribution tails and guarantees that the resulting ECDF predictions to comply with the properties of cumulative distribution functions.

To assess the quality of the resource-occupancy distribution approximation and prediction, we consider the corresponding quantile values. This allows for an intuitive interpretation of the approximation and prediction error.

Artificial neural networks

Artificial neural networks (ANNs) capture nonlinear relationships between input and outputs through a more general and flexible functional representation than traditional statistical methods and are commonly used for regression [7].

Elements and structure of ANNs A brief introduction to ANNs and the specific networks applied in our prediction approach will be provided

below. For a more detailed description the interested reader is referred to [7] or [82].

An ANN is composed of an number of nodes, or neurons, that are connected by links. Each link has a numeric weight associated with it. Some of the neurons are connected to the external environment and are designated as input or output neurons. During learning the weights are modified in order to align the network's input and output behavior with the input and targets provided by the environment. Thus, a neuron has a set of input and output links from and to other neurons. Each neuron performs a local computation given the weighted sum of its inputs using a transfer function. Examples for transfer functions are sigmoid or linear functions.

There is a variety of network structures. The main distinction is between feed-forward and recurrent networks. In feed-forward ANNs, links are unidirectional and there are no cycles where in recurrent networks any link structure is possible. In this chapter we will focus on feed-forward ANNs as they are well understood [7].

In an ANN, neurons are typically arranged in layers which means that each neuron is only linked to neurons in the next layer and that there are no links between neurons of the same layer or links to skip a layer. A further distinction is to be made between input, output and hidden layers. Figure 3.16 illustrates an example for a feed-forward neural network with one hidden layer. The neurons belonging to the input layer compute their output directly from the network's input. The output of the output layer can be the parameters of a class of distribution functions to be predicted.

Models for ANNs employed in this chapter In this chapter we study three regular models for ANNs that are widely-used in applications. Specifically, we consider feed-forward multi-layer perceptrons (MLPs), radial basis function networks (RBNs) and generalized regression neural networks (GRNNs) [7].

The nodes of MLPs each compute the biased weighted sum of their inputs which is used as input for the transfer function to determine the nodes' output. The nodes are arranged in a layered feed-forward structure. The network thus has a simple interpretation as a form of input-output model. MLPs can model functions of almost arbitrary complexity to any desired accuracy given an infinite number of neurons and hidden layers [7]. Important issues in MLP design include determining the number of hidden layers and the number of neurons in these layers [7]. For training the network, i.e. setting the values of the neurons' weights and bias to minimize the er-

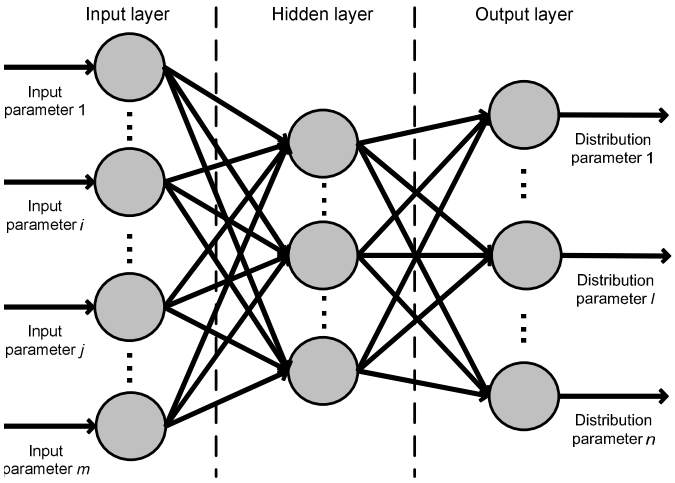


Figure 3.16: A fully-connected feed-forward neural network architecture with one hidden layer and multiple outputs

ror made by the network, example inputs are run through the network and the network's output is compared to the example output. The best-known learning algorithm for MLPs is back-propagation that updates the weights by recursively dividing the observed error among the contributing neurons of the preceding layer and adjusting the weights [7].

RBNs consist of two layers: a hidden radial basis layer and a linear output layer. Since radial basis functions are nonlinear, one hidden layer can be used to model any nonlinear function if the hidden layer contains sufficient radial basis nodes [7]. The linear output layer returns a weighted sum of the radial basis function outcomes. As RBNs can model any nonlinear function using a single hidden layer some design-decisions about numbers of layers can be omitted. The linear transformation in the output layer can be optimized fully using traditional and fast linear modeling techniques [7]. RBNs can thus be trained very quickly: first, the centers and deviations of the radial basis neurons must be set, then the linear output layer is optimized. However, RBNs tend to require many times more neurons than comparable MLPs. Therefore, RBNs tend to be slower to execute and consume more space than MLPs [7].

GRNNs are a type of neural networks that use a kernel-based approach [86] to approximate the underlying probability distribution function of the inputs. The GRNN architecture comprises a radial basis layer and a special linear output layer. The output is determined using a weighted average of the outputs of the training cases, where the weighting is related

to the distance of the input case from the training cases so that cases that are "nearby" contribute most heavily to the calculation. GRNNs are often used for function approximation and have the advantage to train almost instantly. However, since the networks actually contain the entire set of training cases, the execution of GRNNs is space-consuming and slow. For a more detailed description of GRNNs and the concept of kernel-based density estimation the interested reader is referred to [7] and [86].

Distribution approximation

In order to reduce the complexity of supervised learning of an ECDF, we approximate the ECDF by an appropriate class of probability distributions and learn the corresponding distribution parameters.

We will approximate the ECDF by a Gaussian mixture (GM) distribution as the GM distribution can represent a wide range of different distribution shapes that are present in the simulation data as described in Section 3.4.3.

A GM distribution is defined as the convex sum of $k \geq 1$ Gaussian distributions. The cumulative distribution function is given by

$$F_{GM}(y; k, \theta) = \sum_{l=1}^{k} \alpha_l \cdot \phi_{\mu_l, \sigma_l^2}(y),\ y \in \mathbb{N}, \tag{3.7}$$

with $\theta = (\alpha_l, \mu_l, \sigma_l,\ l \in \{1, \ldots, k\})$. ϕ_{μ_l, σ_l^2} denotes a Gaussian cumulative distribution function with mean μ_l and variance σ_l^2 and α_l denote the convex mixing coefficients with $\sum_{l=1}^{k} \alpha_l = 1$. The parameter estimates of θ, $\hat{\theta}$, are determined using the Expectation-Maximization algorithm [30].

Model selection We select the model that minimizes the absolute error between approximated and measured empirical resource-usage quantiles averaged over all the data samples, i.e. the mean absolute error (MAE) of the distribution approximation, given by

$$MAE(k, \hat{\theta}) = \frac{1}{\#\text{samples}} \sum |\hat{F}^{-1;u}_{t_j;\mathbf{a}_{[t_i,t_{i+1}]}}(q) - F^{-1}_{GM}(q; k, \hat{\theta})|.$$

Consider for example the data and GM distributions represented in Figure 3.17. Here, the histogram depicts the occupancy data for a certain day at the CTS-ward with the fitted GM density functions for $k = 1, \ldots, 4$ depicted by the marked lines. For a robust analysis, we consider multiple quantiles

between 70% and 95%. The corresponding quantile values are given in Table 3.5. The visually more appropriate fit obtained for $k = 4$ components of the GM distribution results in the same quantile estimations as the $k = 2, 3$ fit with the $k = 1$ model consistently overestimating the measured quantile values due to the larger spread. The $k \geq 2$ models overestimate the 70% and 85% quantiles but provide an exact estimation for the remaining quantile values. In this situation, the GM distribution with $k = 2$ would be chosen as data representation since the quantile fit equals the estimates obtained by GM approximations with more components while the number of distribution parameters is reduced to 5 parameters versus 8 and 11 parameters for $k = 3$ and 4, respectively.

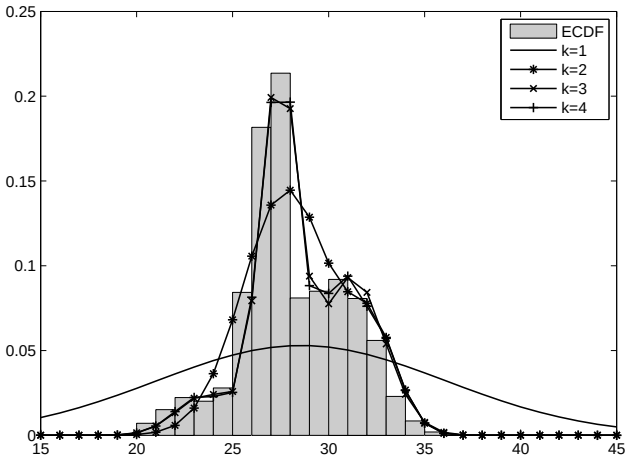


Figure 3.17: Histogram of a example ECDF for CTS-ward with fitted GM distributions with $k = 1, \ldots, 4$ components

Method	$\mathbf{q = 70\%}$	$\mathbf{q = 75\%}$	$\mathbf{q = 80\%}$	$\mathbf{q = 85\%}$	$\mathbf{q = 90\%}$	$\mathbf{q = 95\%}$
ECDF	32	33	33	33	34	34
GMD with $k = 1$	34	35	36	37	38	40
GMD with $k = 2$	33	33	33	34	34	34
GMD with $k = 3$	33	33	33	34	34	34
GMD with $k = 4$	33	33	33	34	34	34

Table 3.5: q-quantiles determined by ECDF and fitted GM distributions with $k = 1, \ldots, 4$ for the example depicted in Figure 3.17

3.5.2 Input features

As explained above, we use a separate ANN for the different units in the care unit network to predict the parameters of the future resource-usage probability distribution. The input of the prediction is composed of the system's state and the admission scheme. Similar to the forward simulations described in Section 3.4, the system state is determined by the patients currently admitted to the different units and their patient groups. Specifically, we use aggregated patient admission information at the corresponding unit and directly preceding units according to the relevant patient flows. The latter was included in order to account for possible future patient flow. The employed input information is summarized in Table 3.6 for the different units.

The time that the patients have been admitted to the current unit is not included in the unit state description as preliminary experiments have shown that the ANNs had difficulties in finding a smooth functional relation between state & admission scheme and the resource-occupancy distribution parameters due to the large possible state space.

Unit	**State information**
CTS-HC	no. type I & IV patients admitted to CTS-HC at t_i
IC	OR scheme for type II patients at t_i, no. type I patients admitted to CTS-HC at t_i, no. type I, II, III, IV patients admitted to IC at t_i
IC-HC	no. type I patients admitted to CTS-HC at t_i, no. type I, IV patients admitted to IC-HC at t_i
MC	OR scheme for type II patients at t_i, no. type I patients admitted to CTS-HC at t_i, no. type I, II patients admitted to MC at t_i
CTS-ward	no. preoperative type I and II patients admitted to CTS-ward at t_i, no. postoperative type I and II patients admitted to CTS-ward at t_i, OR scheme for type I, II patients at t_i, no. type I patients admitted to CTS-HC at t_i, no. type I, II patients admitted to IC at t_i, no. type I patients admitted to IC-HC at t_i, no. type I, II patients admitted to MC at t_i

Table 3.6: State information supervised learning for different units $h = 1$

Since the admission scheme for unconstrained admission control given in Table 3.1 remains unaltered throughout a simulation run, the inclusion of the admission scheme did not yield any performance improvements of the trained ANNs and was therefore omitted during training and evaluation.

3.5.3 Experimental evaluation

The data used for the training and evaluation of the supervised learning approach presented in this section is obtained using the simulation setup described in Section 3.4.3. The admission schemes are determined by Table 3.1 and Algorithm 1 for (un-) constrained admission control.

Setup of GM approximation

Based on the data samples obtained from forward simulation as described in Section 3.4.3 the parameter estimates, $\hat{\theta}$, of the GM distribution function are determined for $k \in \{1, \ldots, 4\}$ using the Expectation-Maximization algorithm [30] that is implemented in the Matlab Statistics toolbox v7.0. As explained in Section 3.5.1, the final model selection is based on the mean absolute error between the measured and approximated quantile values, $MAE(k, \hat{\theta})$.

Setup of ANNs

As explained in Section 3.5.1 our goal is to predict the GM parameters of the future resource-occupancy distribution given the state and (where applicable) information on planned patient admissions. Denote the predicted distribution parameters by $\tilde{\theta}$. The ANN models were created, trained and evaluated using the Neural network toolbox implemented in Matlab v6.0.1. Specifically, the MLPs were built with two hidden sigmoid layer with each 30 neurons and a positive linear output layer. The mean squared error was used as performance measure for training the ANNs, defined as the average squared error between the network's output, $\tilde{\theta}$, and the target values $\hat{\theta}$ over all the example pairs, i.e.

$$MSE(\tilde{\theta}) = \frac{1}{\#\text{samples}} \sum (\hat{\theta} - \tilde{\theta})^2.$$

The MLPs were trained using the Levenberg-Marquardt algorithm, a variant of the back-propagation algorithm, for up to 100 epochs and the deviation of the radial basis neurons of the RBNs and GRNNs was set to 0.8 in order to achieve a close data fit. The RBNs and GRNNs have been designed to

have as many hidden neurons as there are input vectors. The size of the ANNs, as well as the error goals and deviation of the radial basis functions were determined after preliminary experiments considering also constraints imposed by the available memory of the Intel Pentium 4 2.4GHz machine on which the experiments were performed.

As the patient flow at the CTS-PACU and CTS-OR is fully determined by the admission scheme, the prediction for these units is straightforward. For the remaining units, i.e. IC, IC-HC, MC, CTS-HC and CTS-ward, an ANN is to be learnt for each unit separately.

Measuring performance

We evaluate the learned ANNs using the mean absolute difference between the q-quantiles obtained from the predicted GM model parameters, $\tilde{\theta}$, and the empirical q-quantiles obtained from simulation data. Since the actual quantile value to be used for decision support will depend on the specific problem setting and the goal of the hospital management, we evaluated the prediction methods for different quantile values ranging from 0.75 to 0.95 in steps of 0.05. The error, $MAE(q, \tilde{\theta})$, is defined as

$$MAE(q, \tilde{\theta}) = \frac{1}{\#\text{samples}} \sum |\hat{F}^{-1;u}_{t_j;\mathbf{a}_{[t_i,t_{i+1}]}}(q) - F^{-1}_{GM}(q; k, \tilde{\theta})|.$$

Benchmark prediction method

To assess the performance of the ANN prediction method described above, we compared the predictions of the ANNs with the forecasts obtained from a basic prediction heuristic that was inspired by the case study. The benchmark heuristic is based on the resource allocation, r_u, at unit u and determines the forecasts, $\hat{x}^u(q)$, for the q-quantile by

$$\hat{x}^u(q) = \lfloor q \cdot r_u \rfloor \, \forall i, \, h > 0. \tag{3.8}$$

This heuristic was chosen especially with regard to small units where the range of possible occupancy values is small.

Additionally, we considered multivariate linear regression models as benchmark prediction methods. However, the obtained predictions did not result in feasible GM distribution parameters, i.e. positive variances σ_l^2 and mixing coefficients $\alpha_l \in [0, 1]$, cf. Section 3.5.1.

Results

Below we present the main findings obtained from analyzing the results of the distribution approximation and the different prediction methods considered in this section. The numerical results for the GM distribution fit are given in appendix A, Section A.1, Table A.1 and Table A.2. The numerical prediction results obtained using 10-fold cross-validation can be found in Table A.3 and Table A.4.

Distribution approximation Based on the $MAE(k, \hat{\theta})$-based selection method, the overall minimal MAE in the unconstrained admission control case was obtained by a one-component GM distribution for the IC-HC data and two-component GM distributions for the IC, MC, CTS-HC and CTS-ward occupancy data. For the constrained admission control data a one-component model best fit the IC-HC and MC data while two-component GM distributions were chosen for the IC, CTS-HC and CTS-ward.

Fitting a GM distribution to the occupancy data of small units, i.e. MC, IC-HC and CTS-HC, yields low-quality approximations, especially for the 95%-quantile, which is due to the discrete possible occupancy values and the absence of distribution tails in the ECDF. We also fitted Poisson, Gamma, Weibull and uniform probability distributions to the data which resulted in similar or larger $MAE(k, \hat{\theta})$ values compared to the best GMD fit, except for the Poisson distribution which provided better estimates for small quantile values ($q \leq 0.75$) than the GMD fit. However, ANN predictions for Poisson parameters did not appear to improve the results obtained using predicted GM distribution parameters. Therefore, GM distributions were chosen also for MC, IC-HC and CTS-HC occupancy data.

Supervised learning

Basic evaluation This evaluation is based on forward simulation data obtained for the basic allocation at the case study hospital, cf. Section 2.3.5.

In the unconstrained admission control setting, the performance of the different ANNs differs per unit and quantile value. In the majority of cases, GRNNs perform best, with a prediction that differs less than 1 beds for all the units considered, with a small standard deviation of the prediction performance of below 5% with the second best performance obtained from RBNs. For the IC, RBNs outperform the other prediction methods for quantile values up to 0.8. For higher quantile values GRNNs perform (slightly)

better than RBNs, followed by MLPs and the benchmark heuristic with considerable difference in MAE. The different ANNs differ between 0.6 and 2.6 beds in their forecasting error, whereas the benchmark results in an error of more than 4 beds compared to the measured quantile value. Considering the CTS-ward data, GRNNs consistently outperform the other ANNs and benchmark heuristic. The error ranges between 1.3 to 7.6 beds between the different prediction methods. For the MC data, GRNNs yield the best predictions for the 0.7 and 0.85 and higher quantiles, for which the benchmark prediction method performs somewhat better. For the IC-HC, MLPs provide the best forecast for the 70% quantile, GRNNs perform best for the 75% to 85% quantiles and RBNs perform best for the 90% and higher quantile values although the difference between the different ANN predictions is small with at most 0.45. A high deviation between prediction and measured quantile is noticeable for the 95% quantile of 0.85 which amounts to more than 20% of the resource allocation. This deviation, however, can also be explained by the less accurate fit of the GM distribution with one component. For the CTS-HC, the RBNs perform best for the quantiles of 0.85 and smaller and the GRNNs perform best for the larger quantile values.

Overall, MLPs result in the worst prediction performance among the ANNs. This suggests that on the one hand this class of ANNs may not able to approximate the fluctuations in resource occupancy in a stochastic and dynamic problem setting as considered here. On the other hand, the back-propagation learning algorithm may not be able to set the weights and biases of this class of ANNs well enough to obtain good predictions. RBNs yield the best predictions for the CTS-HC and comparatively good predictions for the IC, the IC-HC and MC, but result in poor predictions for the CTS-ward with an error of consistently of more than one bed. In the case of unconstrained admission control the benchmark heuristic performs well for small units due to the discreteness of the possible occupancy values, but performs poorly for the units with more beds.

The overall prediction performance can potentially be improved by ensembling different ANNs for different quantile values of interest to the hospital management. Especially for the IC, MC, CTS-HC and IC-HC, a naive ensemble that consists of the best ANN for the corresponding quantile yields a maximal relative error of 7.8%, 16%, 1.8% and 15.9%[2], respectively. Compared to a minimal prediction error of 25% for a single case if the prediction deviates from the actual measured occupancy quantile by one bed, for the small units CTS-HC, IC-HC and MC, the standard ANNs consid-

[2]Leaving the 95% quantile out of consideration.

ered in this chapter show good potential for improved resource occupancy prediction. Overall, the performance of the ANN predictions decrease for decreasing unit size.

For constrained admission control, the GRNNs show the best prediction performance among the ANNs with a very small standard deviation of performance for all units considered. For the IC, the RBNs perform second best followed by MLPs and the benchmark heuristic for all quantile values. The different ANN models differ an error of about 2 beds. The same holds for the CTS-ward data, with a considerable difference in the prediction performance among the different methods: the ANNs differ at most about 4 beds, the benchmark produces forecasts that differ by about 9 beds from the measured occupancy quantiles. For the MC data, the benchmark performs slightly better than RBNs for the 70% quantile and outperforms MLPs for quantile values of 85% and higher. For the IC-HC, the benchmark heuristic performs second best for all quantile values. The difference in performance of the different methods for MC and IC-HC data amounts to about 0.5 and 1 bed, respectively. For the CTS-HC data, the RBNs outperform the MLPs consistently, however, the MLPs are for some quantile values inferior to the benchmark predictions.

Since the GRNNs outperform the other prediction methods for all units and quantile values an ensemble of ANNs does not improve the prediction performance.

Relative to the resource allocation, the GRNNs result in a maximal relative deviation from actual occupancy quantiles of less than 20%, i.e. 2.4%, 5.4%, 13.5%, 19.9% and 18.7%[2] for the CTS-ward, IC, MC, CTS-HC and IC-HC, respectively. Again, the performance decreases with decreasing unit size.

Visual inspection of the predicted quantile values for the different data sets reveals a considerable variability in resource occupancy, especially for the constrained admission control data with the different admission schemes. In Figure 3.18 some example samples for the CTS-ward data and the corresponding measured and predicted 90%-quantile values are depicted. The quantile values determined on the basis of the occupancy samples are depicted by the black solid line with a large range of fluctuation. The GRNN predictions follow the measured quantile fairly closely with partly more excessive fluctuation than the predicted values. RBN predictions fluctuate less and the MLP predictions remain almost constant with some upward deflections. Due to the variability of the true ECDF the allocation-based

benchmark heuristic renders predictions that consistently differ from the exact occupancy by more than one bed.

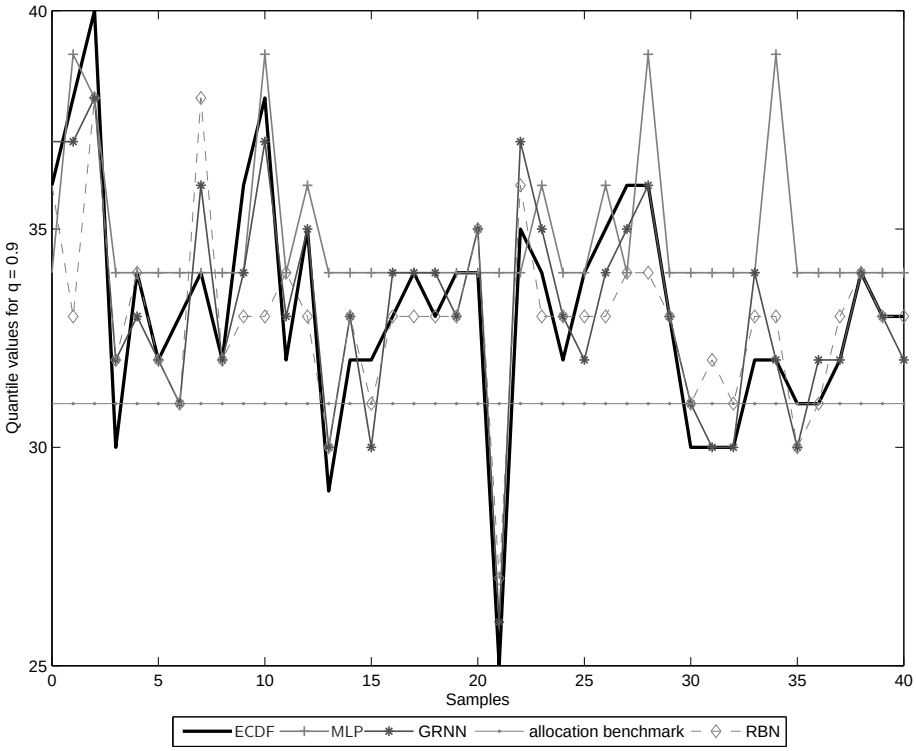


Figure 3.18: Example for predicted 90%-quantile values obtained from the different ANNs and benchmark method for the CTS-ward data

Sensitivity analysis and generalization In a sensitivity analysis we evaluate the performance of the ANN prediction approach for varying resource allocation through (non-)linear scaling of the basis allocation with a percentage of 20%. The linearly scaled allocations are obtained using (3.2) and the non-linear scaling determined by reversely scaling the IC and CTS-ward capacity by $\pm 20\%$ and $\mp 20\%$, respectively. In order to generalize the prediction ANNs, the input features were extended to include the allocation information, r_u, $u \in U$, cf. Table 3.6. The numerical results are given in Table A.5 and Table A.6 in appendix A.1.

In the unconstrained admission control setting, the overall performance decreases compared to the results for the basic resource allocation. Two

effects are to be seen: for the CTS-HC, IC-HC and MC, the prediction performance remains comparable with moderate decrease for some quantile predictions where the allocation is varied only by linear scaling. For IC and CTS-ward, however, the loss in prediction performance is somewhat greater for the smaller quantile values but remains comparable for the larger quantile values ($q \geq 0.85$).

Also under varying resource allocations, a naive ensemble of GRNNs and RBNs can improve the prediction performance of the ANN-based approach. Specifically, an ensemble of GRNNs and RBNs for $q = 0.7, \geq 0.85$ and $q = 0.75, 0.8$, respectively, for the IC data results in a maximal deviation of 0.8 beds and a relative error of 7,3%. Moreover, an ensemble of benchmark heuristic (for $q = 0.75, 0.8$), RBNs (for $q = 0.85$) and GRNNs (for $q = 0.7$ and $q \geq 0.9$), provide for a slight increase in prediction performance of about 2% for the MC. For the CTS-HC, an ensemble of RBNs (for $q \leq 0.85$) and GRNNs achieves a maximal prediction improvement of 13,5%.

For constrained admission control, varying the resource allocation also decreases the prediction performance for the resource occupancy. In general, the performance decrease is greater for the MLPs, RBNs and benchmark heuristic than for GRNNs, which performed best among the ANNs for all units. For the CTS-HC and IC-HC data, the prediction performance of GRNNs remains comparable. The largest decrease in performance occurs for GRNNs and the IC data with on average 42%, however the relative error of GRNNs remains below 10% of the mean number of allocated resources. The decrease of performance of the best predictor, the GRNNs, for the MC and CTS-ward data is limited to about 6%, respectively.

In conclusion, the ANN predictions appear to provide robust and accurate predictions for the occupancy quantiles for the (un-)constrained admission control settings, respectively, also under (non-)linearly scaled resource allocations. Moreover, the predictions in the unconstrained admission control case can be further improved through naive ensembling of different ANNs depending on the quantile values and unit data.

3.6 Conclusions

In this chapter we investigated two approaches for predicting future hospital resource occupancy: forward simulation and supervised learning using artificial neural networks (ANNs). Since the resource occupancy at a hospital unit behaves like a stochastic process, we model the daily resource usage

as a probability distribution and base the forecasts on the current resource occupancy and the planned patient admissions for one or more days in the future. To the best of our knowledge, this is the first simulation-based prediction approach for hospital resource-occupancy distributions. Due to the complex and stochastic patient pathways and the actual patient flow being the result of resource availability, the occupancy distribution is typically unknown and we determine the distribution by maximum likelihood estimation of the empirical distribution function. Furthermore, we showed that the empirical distribution polynomially converges for an increasing number of samples and determined the number of samples required for accurate and precise resource-occupancy distribution predictions.

In order to reduce the computational effort associated with forward simulation we evaluated whether ANNs could be used to approximate the resource-usage distribution. We considered three regular neural network models that were trained using sample data obtained from forward simulation. Our numerical results showed the potential of the ANN prediction approach for this setting. Given a minimal relative error of 25% if the prediction deviates from the observed value by 1 bed for one instance, the maximal relative error of at most 20% of the resource allocation obtained by our approach is promising. The best performing ANN, the GRNNs, resulted in a mean absolute prediction error of at most 1 bed. In a stochastic and dynamic problem domain as the hospital setting considered in this thesis, these results support the applicability of the approach and shows the potential for decision support. Furthermore, we have shown that the ANNs can outperform a basic benchmark prediction heuristic. Moreover, the trained prediction models provide good predictions under varying resource allocations, provided the allocation information is included as input feature. This generalization not only demonstrated the robustness of the proposed prediction methods, but also showed the feasibility of the prediction approach in situations where the resource allocation varies over time. The performance of the ANN distribution predictions could be further improved by naive ensembling of different ANNs depending on the quantile value and unit data. Further research on ANNs is needed to extend the ANN-based prediction approach to situations with greater fluctuations of resource allocation. This, however, is beyond the scope of this thesis.

The extremely hard interdependencies of the underlying stochastic processes, i.e. the stochastic patient pathways, admission and transfer processes, require large networks with the ANN prediction approach studied in this chapter. Consequently, considerable learning data and learning time is required which induces a trade-off between the ANNs' usefulness and the com-

putational efficiency of this approach. In the remainder of this thesis, we will apply forward simulation to an effectual extent to enhance decision-making in resource management discussed in the subsequent chapters of this thesis, albeit at the cost of additional computation time.

Moreover, we extensively analyzed the resource-occupancy distributions obtained from forward simulation. Based on our analysis we could conclude that the normality assumption sometimes assumed in the operations research literature often does not hold statistically in the system under examination. However, the effect of the normality assumption on the hospital operations will depend on the control variables under consideration.

Whereas in this thesis we use prediction information for patient admission control and hospital resource management management, predicted resource-occupancy distributions could also provide valuable decision support for hospital management that is beyond the scope of this thesis. For example, the predictions can indicate the risk of future resource shortage based on which ambulance services or general practitioners could be alerted to send patients to other hospitals for the critical period of time. Moreover, hospital managers could prepare for potential peak capacity, e.g. by exploring possibilities of additional staffing beforehand to be available if necessary. The versatile application is also due to the flexible modeling approach taken in this chapter which allows the calculation of manifold statistics to be used for decision support.

Chapter 4

Multi-objective optimization for hospital resource management

Allocating resources to hospital units is a major managerial issue as the relationship between resources, utilization and patient flow of different patient groups is complex. Furthermore, the problem is dynamic as patient arrivals and pathways are stochastic. In this chapter we present an approach for simultaneously optimizing the allocation of multiple types of resources to different hospital units taking multiple conflicting objectives into account. Specifically, we apply a multi-objective evolutionary algorithm (MOEA) for the optimization. We demonstrate the applicability of this approach to a real-world problem setting and show that the optimized allocations effectively improve resource allocations used in hospital practice. Parts of this chapter have been published in [48].

4.1 Introduction

Hospital resource management is concerned with the efficient and effective allocation of hospital resources, i.e. operating rooms and beds, when and where they are needed. Matching the resource allocation to the demand for care is a stochastic problem as resource usage at a hospital unit behaves like a stochastic process. This is caused by emergency patients arriving in urgent need for care, acute complications of admitted patients that require unexpected patients' transfers and stochastic patients' length of stays (LoS). Furthermore, multiple patient pathways need to be considered when

assigning resource capacity to a hospital care unit as pathways often share resources, e.g. the Intensive Care unit (ICU). Thus, hospital resource management is a complex and highly dynamic problem.

Goals for the optimization of hospital resource management are amongst others a high patient throughput, i.e. the number of patients discharged from the hospital after treatment, low resource costs, infrequent and short-time usage of back-up capacity and a small number of canceled surgeries. We represent these performance measures in three objective functions that have been shown to be conflicting, cf. Chapter 2, Section 2.4.3. In order to allow hospital management to make an effective trade-off between these different objectives we apply a multi-objective (MO) optimization approach that considers the different objectives simultaneously.

As discussed in Chapter 2, an analytical evaluation of a resource allocation is not feasible due to the stochastic patient pathways and the actual patient flow being the result of resource availability. Furthermore, changing the structure of the patient pathways or the underlying probability distributions is non-trivial in an analytical model. Therefore, the simulation described in Chapter 2 is required to be used for the evaluation of a resource allocation. Moreover, we need to consider each unit in the hospital model as many patient pathways involve multiple units. Therefore, it is essential to coordinate the resource allocations at the different stages in these processes in order to prevent local mismatches between patient flow and available beds, cf. Chapter 2, Section 2.4.3. Thus, the decision space comprises allocations for each unit in a hospital. Generating optimal solutions in large, complex search spaces is usually intractable. Thus, an efficient approximation method, that is able to deal with multiple, competing objectives and large, complex search spaces is required. Among the heuristic methods capable of dealing with large search spaces, multi-objective evolutionary algorithms (MOEAs) are promising candidates as their population-based search is capable of generating a set of solutions in one optimization run and they have been shown to be very powerful for MO optimization [14, 20, 32].

Thus, hospital resource management is a complex and dynamic problem that requires state-of-the-art techniques from MO research. Specifically, we apply the SDR-AVS-$\mathbb{M}$ID$\mathbb{E}$A algorithm [15], a multi-objective estimation-of-distribution (MOEDA) algorithm. In contrast to MOEAs that maintain a population of candidate solutions and use blind variation operators such as crossover and random mutation throughout the search, MOEDAs analyze the population by estimating a probability distribution in the solution space. To stimulate the search for a broad range of optimal solutions SDR-AVS-$\mathbb{M}$ID$\mathbb{E}$A uses mixture distributions to cluster the objective space. Addition-

ally the algorithm contains techniques to prevent premature convergence of the EDA. The results obtained by our approach are threefold: (1) we show that the concurrent optimization of multiple units' allocations and objectives effectively improves current resource allocation practice without large investments in additional capacity or homogenization of the resulting patient mix, (2) we demonstrate the applicability of SDR-AVS-MIDEA in combination with a large-scale simulation using a real-world problem instance obtained from the CHE case study, cf. Section 2.3.5, and (3) we analyze the SDR-AVS-MIDEA settings that provide for good results in reasonable time for this MO problem.

The remainder of this chapter is organized as follows. First, we describe related previous work in Section 4.2. Then, we provide a model for the resource allocation problem in Section 4.3. Next, our approach is presented in Section 4.4. We end this chapter with our conclusions.

4.2 Related work

Work on hospital resource management can be found in the operations research and operations management literature. The work in [60, 99] provide theoretical results for hospital bed utilization which is applied for determining efficient resource allocations. This model is not applicable to our problem setting as decision policies and alternative patient routing is not considered in the mathematical model. Moreover, our approach is more flexible and can easily be adopted to other hospital settings. Multiple studies focus on aggregated bed allocation policies, e.g. [40, 89, 97], or resource allocations at single units, e.g. [55, 80]. Aggregated allocation policies are not suited for decision support on resource management in our problem setting as we are interested in allocations on a unit-level. In addition, our approach allows for an in-depth analysis of resource allocations also on the level of different hospital units. Furthermore, their work solely addresses the allocation of hospital beds whereas we consider the concurrent optimization of different types of hospital beds and concerted OR time slot allocation. Earlier work on single unit allocation optimization problems has focussed specifically on the ICU. Ridge et al. [80] present a simulation model based on a case study that is used for the optimization of the ICU bed allocation and propose a reservation policy for emergency patients. In Kim et al. [55] the issue of pooling beds for different specialties at the ICU is addressed. Single unit optimization approaches are not suited for decision support in this problem setting for which it appeared that the optimization of single unit resource

allocations influences multiple patient flows and may create mismatches between the demand for care and the available resources at other care units, cf. Chapter 2. Therefore, the concurrent allocation of resources to multiple pre- and postoperative care units and OR time slots need to be considered to improve the overall performance. In VanBerkel [93] surgical patient flows are simulated to evaluate different resource allocation decisions at the OR and dedicated care facilities. While their simulation is applied for what-if analysis of the different allocations, we present a optimization approach to determine the resource allocation considering multiple conflicting objectives. In [9], the MO optimization problem is addressed. The model, however, is restricted to deterministic patient treatment processes which is not applicable in our problem setting. In our approach, we consider multiple complex stochastic treatment processes that can be flexibly adjusted to other settings.

4.3 Model for hospital resource management

In this section we present the decision variables and define the performance measures under consideration. Moreover, we briefly introduce multi-objective optimization and a few important definitions.

4.3.1 Decision variables & model parameters

In this chapter we focus on the number of allocated resources, i.e. hospital beds and OR time slots, as free decision variables in the simulation described in Chapter 2. A free decision variable refers to a control variable that impacts the performance of the simulation. Denote by $\mathbf{r}$ the vector containing the resource allocations to the different hospital units u with $\mathbf{r} = (r_u,\, u \in U)$.

The following model parameters, i.e. the variables whose values characterize the problem instance, are of importance for the hospital resource allocation problem:

- the patient pathway parameters, described in Section 2.3.4 and the involved care units $u \in U$;
- the decision policies of the respective care unit agents concerning patient (re-)transfers, described in Section 2.3.3;
- the unit costs, c_u, associated with a resource allocated at hospital unit $u \in U$;

- the values of the lower and upper bounds, r_u^{min}, r_u^{max}, for the resource capacity to be allocated to unit $u \in U$.

The unit costs, c_u, are used to calculate the total costs of a resource allocation $\mathbf{r}$ which is determined by equation (2.2) on page 53. The values of r_u^{min} and r_u^{max} are typically imposed by the layout of a hospital unit, the available equipment, staff and funds.

4.3.2 Performance evaluation

As shown in Section 2.4.3 a trade-off is needed between multiple conflicting objectives to optimize resource allocation in hospitals, i.e. high patient throughput at low resource costs and low back-up capacity usage. Because the objectives behave like stochastic variables depending on the simulated patient flow, the objectives are to be optimized under uncertainty. Due to the model complexity which requires simulation for model evaluation, our approach to solve the hospital resource management problem is to consider different realizations of the simulation for the same model parameter setting and determine the arithmetic mean of the simulation outcomes as performance for that parameter setting. We denote the different mean outcomes resulting from running the simulation applying resource allocation $\mathbf{r}$ by

$G_0(\mathbf{r})$**:** the mean total patient throughput under allocation $\mathbf{r}$ as defined in Chapter 2,

$G_1(\mathbf{r})$**:** the mean total resource costs as defined in Chapter 2, equation (2.2), and

$G_2(\mathbf{r})$**:** the mean total weighted back-up capacity usage under allocation $\mathbf{r}$ where the weighting factors correspond to the cost factors $c_u, u \in U \backslash \{CTS - OR\}$ given in Table 2.4 on page 54.

In the following section we formalize the multi-objective optimization problem for the hospital management problem and present few important multi-objective definitions.

4.3.3 Multi-objective optimization problem

In general, it is difficult to express weights to combine the three objectives in Section 4.3.2 in a single scalar objective function to be optimized. The balancing of the different objectives is potentially difficult due to different views of the parties involved. For example, hospital managers would typically uprate the patient throughput and the resource costs, while hospital staff might

emphasize the back-up capacity usage in their weighting. Given these different views, we propose a multi-objective (MO) optimization approach where the different objectives are optimized simultaneously. It is important to note that MO problems distinguish themselves from single-objective problems in that no single optimal solution exist but a set of alternative solutions that cannot be ordered in terms of their objective function values.

For optimizing resource management, $G_0(\mathbf{r})$ has to be maximized, while $G_1(\mathbf{r})$ and $G_2(\mathbf{r})$ have to be minimized. This is equivalent to minimizing $-G_0(\mathbf{r})$, $G_1(\mathbf{r})$ and $G_2(\mathbf{r})$. An allocation at a unit u has to satisfy the constraint imposed by the upper and lower allocation bounds, r_u^{min} and r_u^{max}. This results in the following MO optimization problem:

$$\min G(\mathbf{r}) = (-G_0(\mathbf{r}), G_1(\mathbf{r}), G_2(\mathbf{r})) \tag{4.1}$$

subject to

$$\forall u \in U \; r_u \in \mathbb{N} \cap [r_u^{min}, r_u^{max}]. \tag{4.2}$$

In the following we will provide some relevant definitions for MO optimization that will be used in the remainder of this chapter.

MO optimization definitions

In the context of MO optimization the following concepts are of relevance:

Pareto dominance A solution $\mathbf{r}$ (Pareto) dominates a solution $\mathbf{r}'$ if $\forall$ objectives $i : G_i(\mathbf{r}) \leq G_i(\mathbf{r}')$ and $\exists$ objective $i : G_i(\mathbf{r}) < G_i(\mathbf{r}')$, where $G_i(\cdot)$ denotes the ith objective of G

Pareto optimal A Pareto optimal solution $\mathbf{r}$ is a solution that cannot be improved in one objective without worsening at least one other objective, i.e. $\nexists$ solution $\mathbf{r}'$ that Pareto dominates $\mathbf{r}$

Pareto optimal set The set P_S of all Pareto optimal solutions, i.e. the set of trade-off optimal solutions

Pareto optimal front The set P_F of all objective function values corresponding to the solutions in P_S, i.e. $P_F = \{G(\mathbf{r}) | \mathbf{r} \in P_S\}$

In Figure 4.1 an example of a Pareto front is depicted where two objective functions, objective$_1$ and objective$_2$, are to be minimized. The feasible solutions to the problem are represented by the circled points with the solutions on the Pareto front being highlighted in dark grey. Points A and B lie on the Pareto front as objective$_1(A) >$ objective$_1(B)$ and

objective$_2(A)$ < objective$_2(B)$. Point C, on the other hand, is not on the Pareto Front because it holds that objective$_1(x)$ < objective$_1(C)$ and objective$_2(x)$ < objective$_2(C)$ for $x = A$ and B.

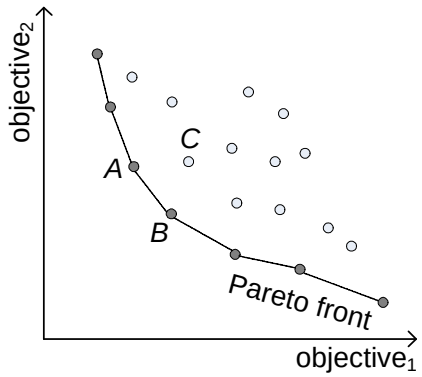


Figure 4.1: Example of a 2-dimensional Pareto front

4.4 Evolutionary multi-objective optimization

In this section, we will present our approach for solving the complex multi-objective optimization problem of allocating resources to a network of care units in a hospital setting. First, we will briefly introduce evolutionary algorithms. Subsequently, we will outline our approach and finally describe the specific algorithm used in this thesis. For a more in-depth introduction on evolutionary algorithms the reader is referred to e.g. Russell and Norvig [82]. A thorough discussion on multi-objective optimization using evolutionary algorithms can be found in Deb [25].

4.4.1 Brief description of evolutionary algorithms

Evolutionary algorithms (EAs) are a search technique that mimics the natural process of biological evolution. The basic idea is that there is a search space that contains some solutions to be found by the EA. The EA starts with distributing multiple solutions into the search space. A solution is also called an individual and the set of solutions at a given moment in time is called a population. The search is an iterative process that generates new individuals from existing individuals of a given population, i.e. variation operators like crossover and/or random mutation are applied to existing in-

dividuals. Every new iteration of an EA is called a generation. Finally, natural selection, i.e. survival of the fittest, determines which individuals of the current population participate in the new population based on quality such that the higher the quality of a solution the higher the chance to survive. Here, the quality of an individual in the population is evaluated by a scalar valued fitness measure. This iterative process is also depicted in Figure 4.2.

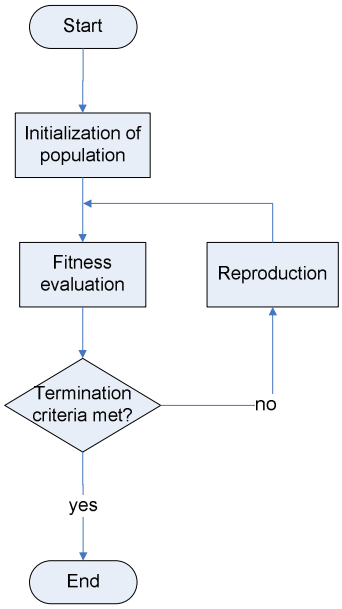


Figure 4.2: Flowchart of an evolutionary algorithm

In the multi-objective (MO) optimization problem considered in this thesis, the EA is applied such that the search space is determined by the parameter space of the optimization. The goal of the EA is then to find the individuals with the best fitness which is determined by the objective functions in the MO optimization problem. It has been found that using EAs is a highly effective way of finding multiple solutions for MO optimization problems due to their population-based approach [25].

4.4.2 Approach

In our optimization approach we apply an estimation-of-distribution algorithm (EDA) which is a class of EAs. Specifically, we use an iterated distribution-estimation evolutionary algorithm (IDEA) for the optimization

of the complex multi-objective resource allocation problem. The main operator of variation in IDEA is the estimation of a probability distribution based on the selected solutions and the subsequent drawing of new solutions from this distribution.

Similarly to EAs, the first population of candidate solutions in an EDA is generated at random and a collection of solutions is selected with better fitness. Then, new candidate solutions are generated by estimating a probability distribution from the selected solutions and by randomly drawing samples from this distribution. The new population is evaluated based on their fitness and the better individuals are kept. This process is iterated in every generation until a predefined termination criterion is met:

1. select a collection of solutions

2. estimate a probability distribution from the selected solutions by learning the parameters of a predefined parameterized representation of the probability distribution

3. generate a collection of new solutions by sampling from the estimated probability distribution

4. replace some of the solutions in the current generation by the new individuals

Using this variation operator, IDEAs aim to induce and exploit the structure of an optimization problem. It has been shown that IDEAs are capable of learning dependencies between problem variables and that they scale up efficiently for some problems where classic EAs do not scale up efficiently [11].

Specifically, we apply an IDEA for MO optimization that integrates an Standard Deviation Ratio (SDR) trigger and Adaptive Variance Scaling (AVS), called SDR-AVS-MIDEA, that was presented in [15]. The algorithm was shown to be an efficient optimization technique for MO optimization problems. A more detailed description of SDR-AVS-MIDEA is given in Section 4.4.3.

The resource allocation vector, $\mathbf{r}$, is optimized using SDR-AVS-MIDEA. The fitness of $\mathbf{r}$ is obtained by applying the corresponding allocation in multiple runs of the agent-based simulation described in Chapter 2, cf. Figure 2.1, and taking the mean of the outcome measures, cf. Section 4.3.2. A graphical representation of this approach is also given in Figure 4.3.

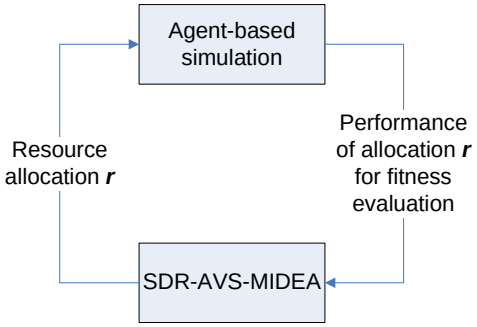


Figure 4.3: Representation of optimization approach using the agent-based simulation described in Chapter 2 and SDR-AVS-𝕄IDEA

4.4.3 Description of SDR-AVS-𝕄IDEA

The SDR-AVS-𝕄IDEA algorithm is summarized in pseudocode in Algorithm 2 [15], and will be described below.

Algorithm 2: Pseudo-code description of the SDR-AVS-𝕄IDEA algorithm

```
1 Initialize population of n (random) solutions
2 Cluster objective space into k clusters of size n^subpop and assign solutions to
  subpopulation k using nearest-neighbor heuristic
3 repeat
4     Select best ⌊τn⌋ solutions by diversity-preserving-selection
5     Generate n − ⌊τn⌋ new solutions:
      estimate a normal distribution for each subpopulation; if SDR triggers
      AVS, apply AVS scheme to scale corresponding covariance matrix and
      draw samples from resulting distribution
6     Replace non-selected solutions with new solutions and assign solutions
      to nearest cluster
7     Compute SDR trigger & AVS scheme for obtained population
8 until termination ;
```

SDR-AVS-𝕄IDEA is a multi-objective IDEA. Specifically, SDR-AVS-𝕄IDEA employs a Gaussian mixture distribution, cf. equation (3.7) on page 103. The mixture distribution is built by clustering the objective space of size into k clusters of size n^{subpop} and estimating a normal distribution for each cluster. The solutions in cluster k are referred to as subpopulation k. In order to stimulate the search for a broad Pareto-front the clusters are kept separated in the objective space throughout the optimization as will be explained below. New solutions per subpopulation are generated according

to the IDEA principle by sampling from the estimated distribution.

The algorithm applies truncation selection, i.e. the best $\lfloor \tau n \rfloor$ solutions are selected to be preserved from one generation to the next. Here, τ denotes the percentile for truncation selection. In order to stimulate a broad Pareto-front, diversity-preserving-selection is applied. The meaning is as follows. For each solution the domination count is determined, i.e. the number solutions that dominate the respective solution. The lower the domination count, the better the solution. If the number of solutions with a domination count of 0 exceeds the $\lfloor \tau n \rfloor$ solutions to be selected, all solutions with domination count equal to 0 are preselected. For the final selection, first a single solution is selected with an extreme value in an arbitrary objective. Then, solutions are added to the final selection by iteratively selecting the solution with the largest distance to the nearest neighbor in the final selection.

To reduce the risk of premature convergence by vanishing variance of the estimated distributions, adaptive variance scaling (AVS) is applied as proposed in [16]. A brief outline of AVS is given below, for a detailed description the reader is referred to [16]. The rationale behind AVS is that the smaller the variance, the smaller the area of exploration for the algorithm. Specifically, SDR-AVS-MIDEA maintains a variance multiplier c^{AVS}. Let Σ denote the covariance matrix of the mixture distribution. Then, upon sampling new solutions the distribution is scaled by c^{AVS}, i.e. the covariance matrix used for sampling from the underlying distribution is $c^{\mathrm{AVS}}\Sigma$ instead of Σ. This means that if the best fitness value improves in one generation, then the current size of the variance allows for progress. Hence, a further enlargement of the variance may allow for further improvement in the next generation. Then, c^{AVS} is scaled by $\eta^{\mathrm{INC}} > 1$. If the best fitness does not improve, the range of exploration may be too large to be effective and the variance multiplier should be decreased by a factor $\eta^{\mathrm{DEC}} \in [0, 1]$.

In addition, standard-deviation ratio (SDR) triggers are used as proposed in [16]. In SDR-AVS-MIDEA a threshold θ^{SDR} is used that triggers the further enlargement of the variance multiplier if the best fitness in a cluster is improved in the last iteration and the average improvement is found to be more than θ^{SDR} standard deviations away from the distribution mean of the corresponding cluster. For a detailed description of the SDR scheme the interested reader is referred to [16].

4.5 Experiments and settings

In this section we describe the experiments that were performed to evaluate the proposed optimization approach. First, we determine the required subpopulation size n^{subpop} and the required number of evaluations to obtain high-quality results at reduced computational costs for the optimization problem at hand. Then, we present the optimization results obtained for the previously determined SDR-AVS-MIDEA settings and analyze the optimized resource allocations and their implications on hospital practice.

4.5.1 Basic algorithmic setup

The settings of the parameters in SDR-AVS-MIDEA are based on the guidelines reported in [17] and the best results reported in [15]. The percentile for truncation selection is set to $\tau = 0.3$, the variance multiplier decreaser of AVS, η^{DEC}, equals 0.9 and $\eta^{\text{INC}} = 1/\eta^{\text{DEC}}$. The SDR threshold is set to $\theta^{\text{SDR}} = 1.0$. As in [15], a so-called "elitist archive" is maintained to retain non-dominated solutions found during optimization. Using the "elitist archive" the objective space is discretized in each objective, here a discretization length of 10^{-3} is used. Then, an individual found during the optimization is added to the set of non-dominated solutions, only if the corresponding cube in the objective space does not yet contain a solution. The employed discretization length was found to provide sufficient granularity for the final Pareto-front approximations [15]. Preliminary experiments showed that $k \geq 4$ clusters are required in order to obtain a broad Pareto-front. Due to the time-consuming fitness evaluation using simulation, a maximum of 1600 generations is allowed for the optimization of different allocations. The number of solutions per subpopulation, n^{subpop}, and the required number of evaluations are determined in Section 4.5.3 taking the convergence of the obtained results into account.

In the EDA representation, an individual corresponds to the parameters of a resource allocation, $\mathbf{r} = (r_u, u \in U)$, i.e. an individual specifies the number of resource allocated to the different units $u \in U$. As explained in Section 4.3.1 the resource allocation at unit $u \in U$ is bounded by r_u^{min} and r_u^{max}. The corresponding parameter values were obtained from domain experts from the CHE. These values are given in Table 4.1.

Measuring performance

Convergence For measuring convergence performance we consider the subset of all non-dominated solutions, called approximation set, that is con-

	CTS-OR	**CTS-HC**	**CTS-PACU**	**IC**	**IC-HC**	**MC**	**CTS-ward**
r_u^{min}	0	0	0	5	2	2	20
r_u^{max}	6	6	6	20	6	10	50

Table 4.1: Resource bounds obtained from CHE case study

tained in the final population that results from running SDR-AVS-𝕄IDℍA and denote this set by $\mathcal{S}$. In running the EA we are interested in finding a good and diverse approximation of the set of Pareto-optimal solutions, P_S. We assess the convergence performance of $\mathcal{S}$ using $D_{P_F \to \mathcal{S}}(\mathcal{S})$ [15] as performance indicator. This performance indicator computes the average distance from over all points in the Pareto-optimal front P_F to the nearest point in a given a approximation set $\mathcal{S}$, i.e. the Pareto front of outcomes obtained from a single run of SDR-AVS-𝕄IDℍA, given by

$$D_{P_F \to \mathcal{S}}(\mathcal{S}) = \frac{1}{|P_S|} \sum_{\mathbf{r}' \in P_S} \min_{\mathbf{r} \in \mathcal{S}} \{d(\mathbf{r}, \mathbf{r}')\}, \tag{4.3}$$

where $d(\mathbf{r}, \mathbf{r}')$ denotes the Euclidean distance between the objective values $G(\mathbf{r})$ and $G(\mathbf{r}')$ of the solutions $\mathbf{r}$ and $\mathbf{r}'$, respectively. The $D_{P_F \to \mathcal{S}}(\mathcal{S})$ indicator represents an intuitive trade-off between the diversity of the approximation set $\mathcal{S}$ and its proximity, i.e. the closeness to the optimal Pareto front P_F. A smaller value of $D_{P_F \to \mathcal{S}}(\mathcal{S})$ is preferable. The minimum of 0 is achieved if and only if the approximation set and the set of Pareto-optimal solutions are identical.

Hospital resource management is a complex real-life optimization problem with a very large search space. Therefore, the set of globally Pareto-optimal solutions, P_S, is typically not known beforehand. For the specific parameter setting used in this chapter, it was possible to calculate the global Pareto front by brute-force optimization using the simulation to evaluate the more than $7.6 \cdot 10^6$ possible resource allocations, determined on the basis of the resource bounds given in Table 4.1. With an average simulation runtime of 1.6 seconds using the simulation settings described in Section 4.5.2, the runtime of the brute-force optimization amounts to about 141 days. The calculation was only made possible due to the exceptional access to a high-performance computing system with more than 500 quad-core nodes running at speeds between 2.26 GHz and 2.5 GHz.

Thus, in general a brute-force approach is infeasible for the problem at hand, especially in a hospital setting where no high-performance computer cluster is available. This especially also holds for analyzing even more com-

plex optimization approaches (as in Chapter 5) or for extended hospital models. Therefore, we use a more generally applicable approach and approximate P_S by 10 independent runs of SDR-AVS-𝕄IDEA with a large number of generations and a large population size. This approach is also applied in Chapter 5 where the set of possible parameter values is uncountable which makes a brute-force optimization approach impossible. A similar approach has also been proposed in [14]. Due to the time-consuming fitness evaluation using simulation, we use a maximal number of generations equal to 1600 and the subpopulation size determined by the guideline in [17], i.e.

$$n^{subpop} \geq 10 \cdot \#\text{parameters}^{0.7} + 10, \tag{4.4}$$

which results in a subpopulation size of 50 for the optimization problem at hand. In Section 4.5.4 we show that the proposed approximation approach yields very good results of SDR-AVS-𝕄IDEA at considerably more tractable computational costs in this setting. Therefore, the $D_{P_F \to \mathcal{S}}(\mathcal{S})$ values in the remainder of this thesis will be computed with respect to the obtained approximation of P_S.

Benchmarking In addition to the convergence, we evaluate the performance of the optimized allocations with respect to benchmark allocations obtained from the CHE case study. Here, we consider the basic allocation, given in Table 2.1 on page 45 and linear variations thereof, denoted by $\mathbf{r}^{CHE \pm i}$ with $\mathbf{r}^{CHE \pm i} = (\lfloor r_u^{CHE} \cdot (1 \pm i) + 0.5 \rfloor,\ u \in U)$ with $i = 10\%, 20\%, 30\%$.

4.5.2 Setup agent-based simulation

To evaluate the quality of a resource allocation we run 10 simulation runs of 16 weeks including 12 weeks of warming-up. Preliminary experiments have indicated that warming-up of 12 weeks were necessary in order to avoid early convergence of the optimization towards minimal allocations due to the empty hospital in the start of a simulation run. The warming-up period is not measured in the simulation outcomes. As we showed in Chapter 2, Section 2.4.2, the simulation outcomes are almost linear in the duration of a simulation run. Therefore, in favor of computationally feasible MO optimization using SDR-AVS-𝕄IDEA in combination with a large-scale simulation a short duration of the simulation can be used and the obtained results will perform strongly also for longer time periods. This setting results in a runtime of about 1.6 seconds for the evaluation of a resource allocation.

During the optimization the simulation uses the same random seeds during execution in order to allow for a fair comparison. In order to assess

whether the solutions obtained from SDR-AVS-𝕄IDEA possibly overfit the resource allocation problem for the fixed seeds, we additionally validated the optimized solutions by evaluating the allocations using 50 different random seeds in the simulation.

4.5.3 Setting the subpopulation size and number of evaluations

Using the subpopulation sizes determined by (4.4) with a maximum of 1600 generations results in a runtime of SDR-AVS-𝕄IDEA of approximately 10 hours. Specifically, we used up to 40 nodes running at speeds between 1.4Ghz and 2.2Ghz. The largest part of this runtime is used for evaluating the fitness of the solutions using the simulation. Although this order of runtime is a substantial reduction compared to the brute-force optimization, it is still infeasible if the proposed optimization approach is to be applied in a hospital setting where typically no high-performance computer cluster is available and the optimization has to be performed regularly on a single PC. Therefore, we study the number of computational resources that is required to obtain solutions that are reasonably close to the Pareto-optimal solutions and show sufficient diversity after a reasonable runtime.

Determining the required subpopulation size

Using (4.4) the subpopulation size is determined as 50 which we varied for our evaluation on the basis of the following considerations. Firstly, (4.4) was determined based on single-objective optimization research and no guideline is available for multi-objective optimization. Moreover, it is unknown whether a mixture distribution over a front in multi-objective optimization can decrease the required subpopulation size since the diversity preserving selection in SDR-AVS-𝕄IDEA and the Pareto front consisting of multiple clusters each containing multiple solutions counteract diversity loss [12]. Therefore, we varied the subpopulation size between 10, 30 and 50. These values were chosen such that a broad spectrum of subpopulation sizes is evaluated.

Using the $D_{P_F \to \mathcal{S}}(\mathcal{S})$ performance indicator, convergence graphs can be computed for the different subpopulation sizes. The average convergence graph shows the value of the $D_{P_F \to \mathcal{S}}(\mathcal{S})$ indicator as a function of the number of generations. In Figure 4.4 such a convergence graph is shown for the allocation problem at hand. For all subpopulation sizes a steep decline of $D_{P_F \to \mathcal{S}}$ can be noted in the first 100 generations, after this the decrease in

$D_{P_F \to \mathcal{S}}$ is reduced. Moreover, the subpopulation size determined by (4.4) achieves a better and faster convergence compared to the smaller population sizes.

The final value of $D_{P_F \to \mathcal{S}}$ differs considerably with a value of about 5 and 7 for $n^{subpop} = 30$ and 10, respectively. For $n^{subpop} = 10$ a non-monotonic decrease can be noted. Here, truncation selection with $\tau = 0.3$ provides for only three solutions to be selected in each cluster that are used to estimate the normal distribution to sample new solutions from. This limited number of solutions slows down the convergence of the solutions to the Pareto-front in this setting.

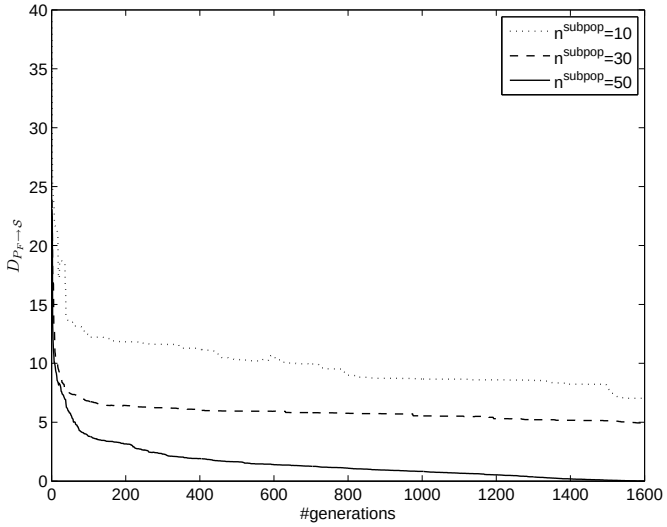


Figure 4.4: Convergence graph for varying subpopulation sizes

Based on these findings the question arises what impact the final $D_{P_F \to \mathcal{S}}$ values for the different subpopulation sizes have on the quality of the obtained solutions with the order of obtained G_0, G_1 and G_2 values being (up to) several hundreds. To address this issue we analyzed the obtained Pareto-fronts for the different population sizes. In Figure 4.5 the results for the optimized resource allocations are presented for the different subpopulation sizes. The Pareto-fronts are depicted with G_1 and G_0 values plotted on the horizontal and vertical axes, respectively, for predefined intervals of G_2 values. In our convergence evaluations, the exact values for back-up capacity usage are of minor importance and a categorization of minimal

(corresponding to $G_2 \in [0, 25)$), very small ($G_2 \in [25, 50)$), small, medium, etc. is therefore sufficient for the representation of the optimization results. Moreover, this two-dimensional representation allows for improved visibility.In the analysis, we confined the results to G_2 values below 200 as a higher back-up capacity usage is not desirable for many hospitals.

The different subpopulation sizes show a noticeable visual difference between the obtained Pareto-fronts for $n^{subpop} = 10$, 30 and 50, especially for $G_2 < 100$. The difference for $n^{subpop} = 30$ and 50, however, seems small. The optimization with a $n^{subpop} = 50$ appears be able to push the obtained front somewhat further towards solutions with higher G_1 value and G_2 value than the optimization with smaller population size. Moreover, it appeared that a larger number of non-dominated solutions and their diversity along the front is achieved for $n^{subpop} = 50$. Thus, $n^{subpop} = 30$ appears to sufficiently push the population into regions with smaller G_1 and G_2 values. For the overall Pareto front, however, the subpopulation size determined by (4.4) appears to be preferable and provides a greater number of trade-off points which is desirable for the decision maker. We observed that the set of non-dominated solutions obtained for $n^{subpop} = 50$ contains about 60% more points. Therefore, depending on the desired range and available number of optimized solutions a smaller n^{subpop} value may be considered for the MO optimization problem at hand.

Determining the required number of evaluations

Since the distance between the Pareto fronts obtained for different n^{subpop} values appears relatively small and the convergence graphs show a slower convergence after an increasing number of generations, also the question arises whether decreasing the number of evaluations may result in comparable fronts to be obtained at reduced computational costs.

To answer this question we can take two approaches:

1. determine the number of generations needed to obtain comparable convergence performance, i.e. the $D_{P_F \to \mathcal{S}}$ value for $n^{subpop} = 50$ and 30 equals the $D_{P_F \to \mathcal{S}}$ value obtained for $n^{subpop} = 10$ value after 1600 generations,

2. determine the number of generations which have the same number of evaluations for $n^{subpop} = 30$ and 50 as for $n^{subpop} = 10$ and 1600 generations, i.e. applying truncation selection, the number of evaluations is determined by

$$\#\text{ evaluations} = k \cdot n^{subpop} + (1 - \tau) \cdot k \cdot n^{subpop} \cdot \#\text{ generations}, \quad (4.5)$$

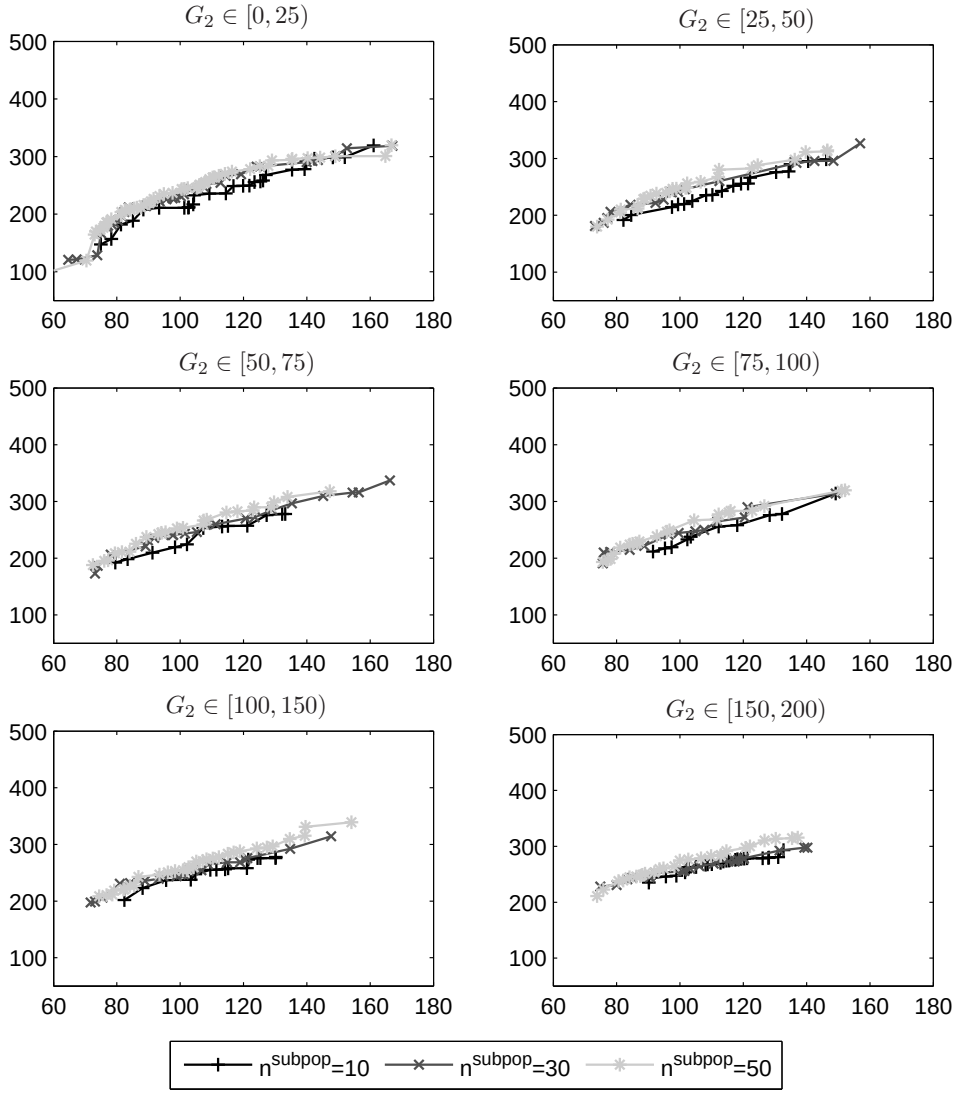


Figure 4.5: Pareto fronts for varying subpopulation sizes after 1600 generations with G_1 and G_0 values depicted on the x- and y-axes, respectively

and we then have to equate the equations for the different population sizes and take the appropriate values of the number of generations for $n^{subpop} = 30$ and 50.

Taking the first approach, the final performance of 7.02 for $n^{subpop} = 10$ is achieved for $n^{subpop} = 30$ and 50 after 80 and 35 generations, respec-

tively. The corresponding Pareto-fronts are shown in Figure 4.6. Compared to the results obtained after 1600 generations for $n^{subpop} = 30$ and 50, the different Pareto-fronts are closer to each other and the range of the Pareto fronts is decreased which constitutes a loss in performance. Thus, the considerably reduced number of generations did not allow the algorithm to push the Pareto-front to a comparable extent. Moreover, the number of available points in the set of non-dominated solutions is considerably reduced (especially for $G_2 > 25$). Based on the obtained Pareto front data, we observed that the number of non-dominated solutions is decreased by a factor of about 2 and 5 compared to the results obtained after 1600 generations for $n^{subpop} = 30$ and 50, respectively. Compared to the results obtained for $n^{subpop} = 10$ after 1600 generations, we noticed that the fronts for $n^{subpop} = 30$ and 50 contain only about half the number of points. Thus, we can conclude that this approach results in a reduced number of generations that is too small for obtaining comparably large and broad Pareto-fronts.

Following the second approach, the number of evaluations involved in running SDR-AVS-MIDEA using $k = 4$ clusters, a selection percentile $\tau = 0.3$ and $n^{subpop} = 10$ for 1600 generations amounts to 44840 which corresponds to about 532 and 319 generations for $n^{subpop} = 30$ and 50, respectively. The results obtained following the second approach are depicted in Figure 4.7. The fronts obtained for $n^{subpop} = 30$ and 50 and reduced number of allowed generations exhibit comparable shapes and locations as for the same subpopulation sizes after 1600 generations. Again, using a subpopulation size of $n^{subpop} = 50$ SDR-AVS-MIDEA appears be able to push the obtained front somewhat further towards solutions with higher G_1 value and G_2 value than the optimization with $n^{subpop} = 30$ and we can observe noticeable differences to the front resulting from $n^{subpop} = 10$ and 1600 generations. Moreover, the number of available points in the sets of non-dominated solutions appear to compare well for the different subpopulation sizes. Compared to $n^{subpop} = 10$, we observed from the data sets that the front obtained for $n^{subpop} = 30$ contains almost the same number of points, for $n^{subpop} = 50$ the size of the set of Pareto-optimal solutions is increased by about 24%. In comparison to the sets of non-dominated solutions obtained for $n^{subpop} = 30$ and 50 after 1600 generations, a decrease of about 19% and 38% can be observed. Therefore, we can conclude that the number of generations needed to obtain comparable optimization results can be reduced by a factor of 5 in this setting. This result enables the multi-objective optimization to be performed on a single PC in about three days. Although a shorter runtime would be preferable this runtime is still reasonable considering the fact that the optimization is solely to be rerun when considerable

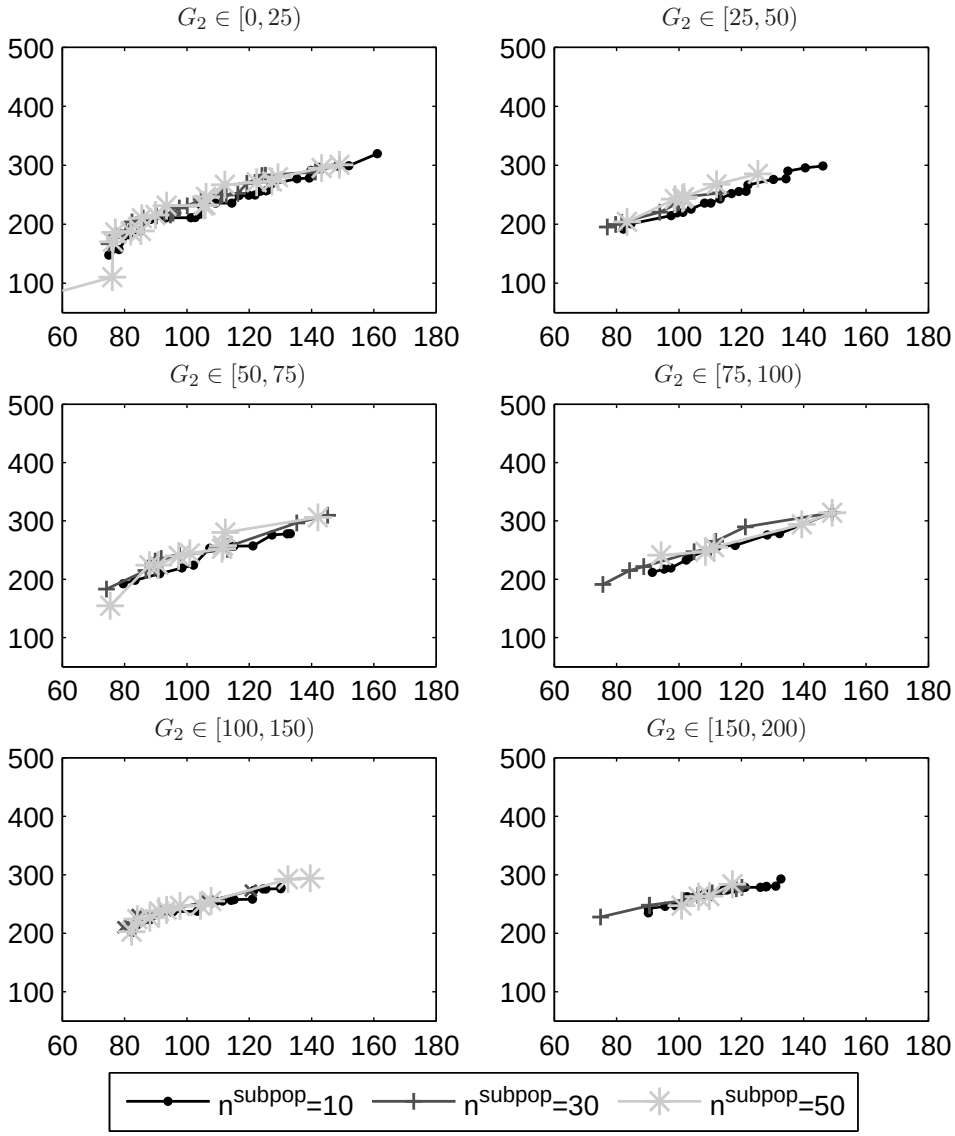


Figure 4.6: Pareto fronts after 1600, 80 and 35 generations for subpopulation sizes of 10, 30 and 50 with equal $D_{P_F \to \mathcal{S}}$ value; the x- and y-axes depict G_1 and G_0 values,

changes in the underlying patient pathways have occurred.

Intermediate conclusions Due to the fast convergence of SDR-AVS-MIDEA for the optimization problem at hand, a considerable reduction in

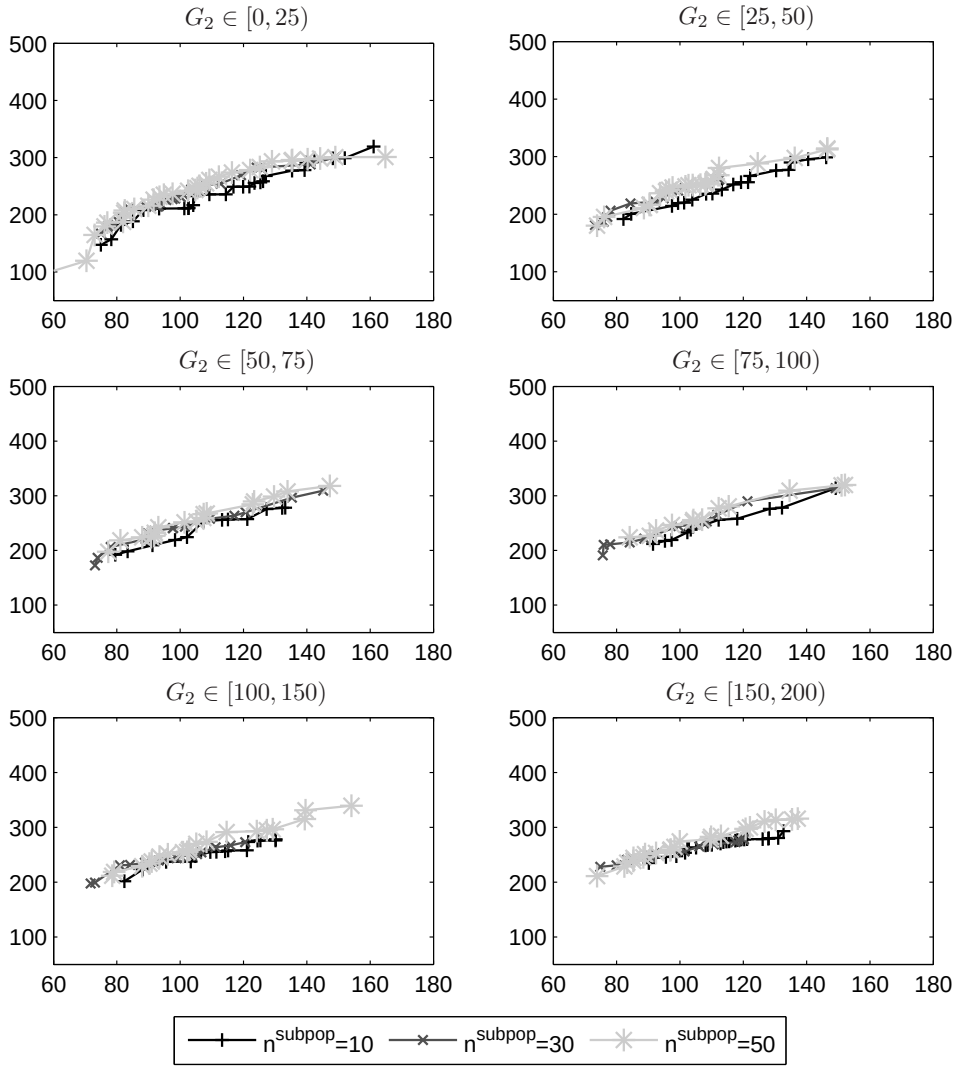


Figure 4.7: Pareto fronts after 1600, 532 and 319 generations for subpopulation sizes of 10, 30 and 50 with equal number of evaluations; the x- and y-axes depict G_1 and G_0 values, respectively

runtime can be achieved by reducing the number of allowed generations for the bigger subpopulation sizes which result in similar Pareto fronts at the expense of minor losses in the number of non-dominated solutions compared to the initial setup of the MO optimization. The resulting runtime allows the optimization to be performed in hospital practice within reasonable time. If

the region of interest to the decision maker is restricted to smaller values of G_1 and G_2, a further considerable reduction in runtime can be obtained by decreasing the subpopulation size which also involves a further loss in the size of the Pareto-optimal sets.

4.5.4 Optimization results

In order to provide a large range of the optimized allocations, the following optimization results are obtained by 10 optimization runs of SDR-AVS-$\mathbb{M}$IDEA using $n^{subpop} = 50$ and a maximum of 532 generations. To crossvalidate our results, we also evaluated the obtained optimized allocations using the simulation with 50 different random seeds. Also, we compare our results to the global Pareto-optimal solutions obtained by brute-force optimization.

Analysis of obtained Pareto fronts

The results in Figure 4.8 show that the benchmarks determined from current hospital practice and variations thereof are dominated by the optimized allocations obtained by our approach. The crossvalidation evaluations are comparable the EDA results obtained for 10 fixed random seeds and also dominate the benchmark allocations. Although the difference in performance between the optimized and benchmark allocations appears rather small in Figure 4.8, the reader should note that the results are based on simulation runs of 4 weeks which means that a small difference achieved in this period is increased considerably when considering the performance on a yearly basis.

In order to provide an overall picture of the optimized resource allocations, a three-dimensional plot of the obtained performance is presented in Figure 4.9. Since the crossvalidation results are comparable to the optimization results obtained for 10 random seeds, solely the optimization results obtained from the EA are shown. The depicted Pareto-front is concave for G_2 values up to 150, for larger G_2 values the front curves into a convex shape. We can see that most of the Pareto-optimal solutions have G_2 values of less than 50 or more than 150 with G_1 values between 60 and 120. For G_2 values between 50 and 250 less solutions are found by SDR-AVS-$\mathbb{M}$IDEA which is to be attributed to the combinatorial nature of the optimization problem. Also, the layers that can be observed for G_2 values up to 50 can be ascribed to the discreteness of the resource management problem. However, solutions with a G_2 value of 200 and higher may not be desirable for many

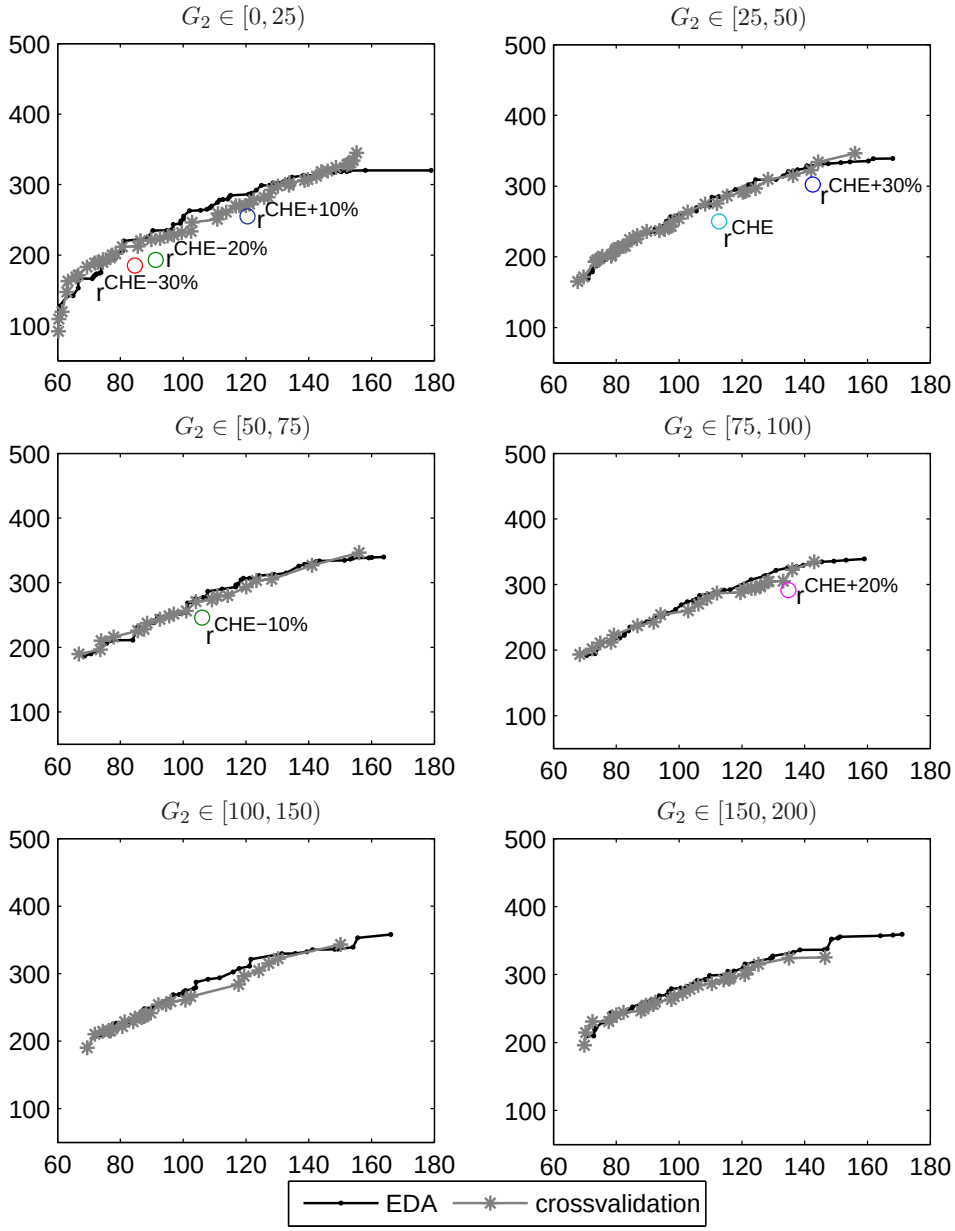


Figure 4.8: Pareto fronts obtained from multiple optimization runs including crossvalidation results and benchmark allocations from CHE case study

hospitals.

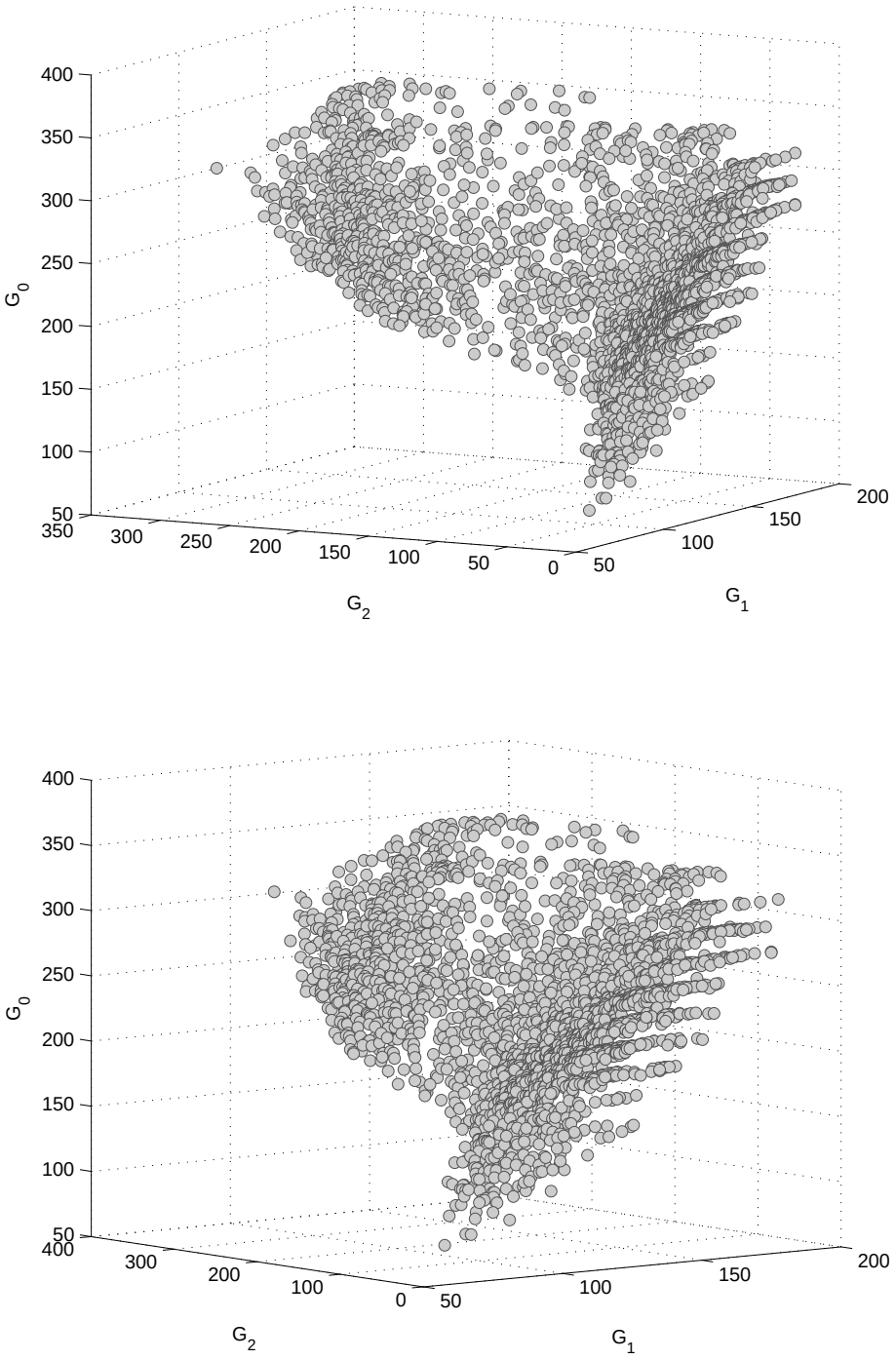


Figure 4.9: Three-dimensional Pareto front obtained from multiple runs of SDR-AVS-𝕄IDEA for $n^{subpop} = 50$ and 532 generations from two perspectives

Comparison to brute-force optimization results In order to provide an overall impression of the globally Pareto-optimal resource allocations, a three-dimensional plot of the obtained performance obtained by brute-force optimization is presented in Figure 4.10. We observe that the depicted Pareto-front is comparable in shape to the Pareto-front obtained by SDR-AVS-𝕄IDEA being concave for G_2 values up to 150 and curving into a convex shape for larger G_2 values. We can also see that the brute-force optimization results in a larger number of available Pareto optimal solutions, especially for G_2 values between 100 and 250 and for extreme G_0, G_1 and G_2 values.

For a better comparison, the two-dimensional Pareto fronts obtained by brute-force optimization and the proposed evolutionary MO optimization using SDR-AVS-𝕄IDEA are depicted in Figure 4.11. We can observe that the Pareto front obtained by the evolutionary MO optimization approach closely approximates the brute-force optimization Pareto front. Only for G_2 values larger than 100 a noticeable difference can be seen. For smaller G_2 values, our approximation yields good to very good results, especially for G_1 values up than 150. In accordance with the observation above, the brute-force Pareto front runs further for very small or large G_1 values. As opposed to the 141 days needed to perform brute-force approach, the MO optimization using SDR-AVS-𝕄IDEA can be run in 3 days on a regular PC which makes our approach practically feasible.

Intermediate conclusions We can conclude that the proposed MO optimization approach efficiently improves the benchmark allocations obtained from current hospital practice. Moreover, the crossvalidation results show the robustness of the obtained results. In comparison to the brute-force optimization results, the optimized allocations obtained by SDR-AVS-𝕄IDEA yield a very good performance, especially for resource costs below about 140 and smaller G_2 values. Furthermore, the SDR-AVS-𝕄IDEA results can be obtained in a fraction of the time needed to perform the brute-force optimization which makes the proposed MO approach practically feasible in a real-life hospital setting.

Analysis of optimal allocations and resulting patient flows

Due to the large number of optimized solutions and corresponding parameter values, we analyze the obtained resource allocations and the resulting patient flows using descriptive summary statistics. Figure 4.12 and Figure 4.13 show boxplots of the optimized allocation parameters and the re-

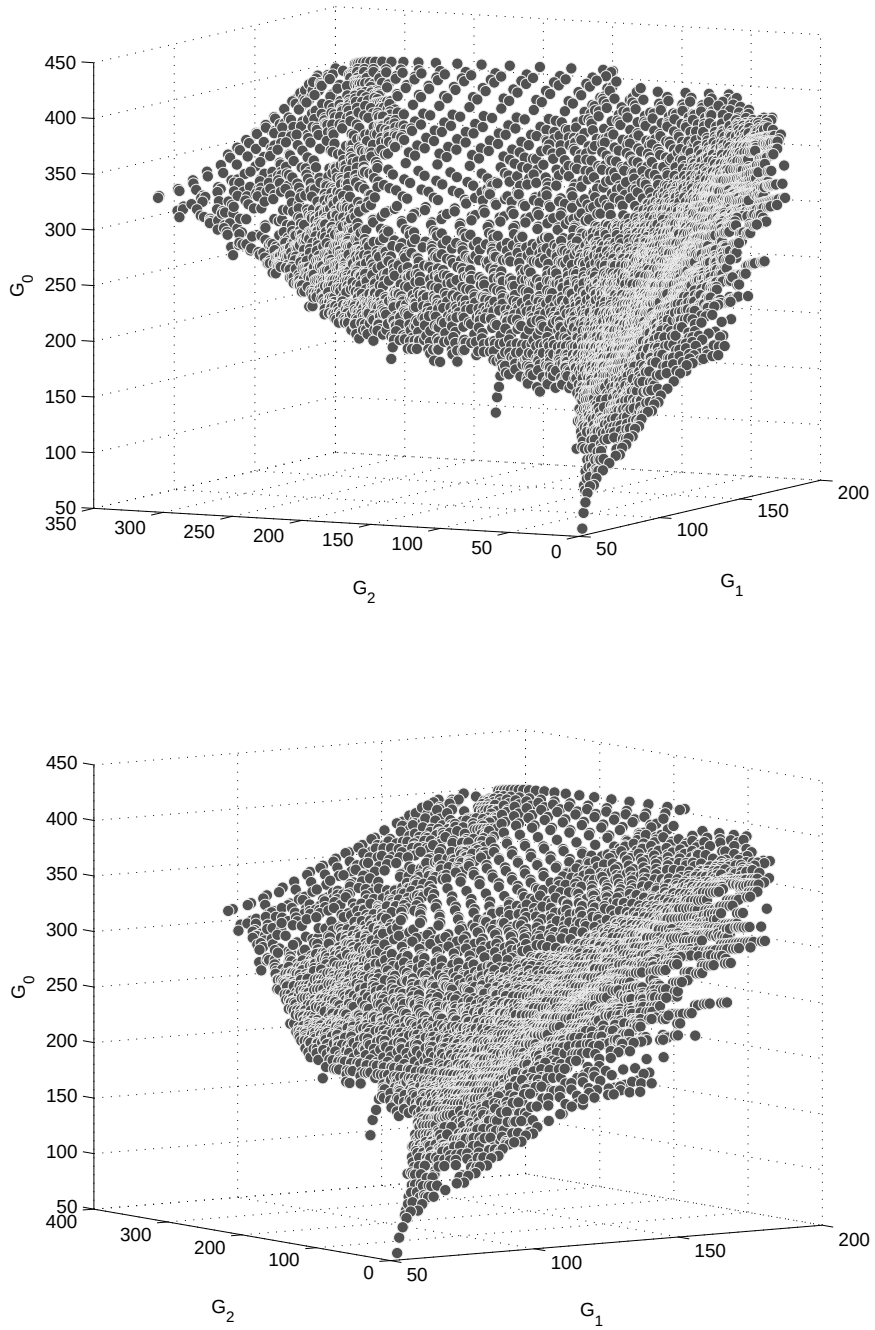


Figure 4.10: Three-dimensional Pareto front obtained from brute-force optimization from two perspectives

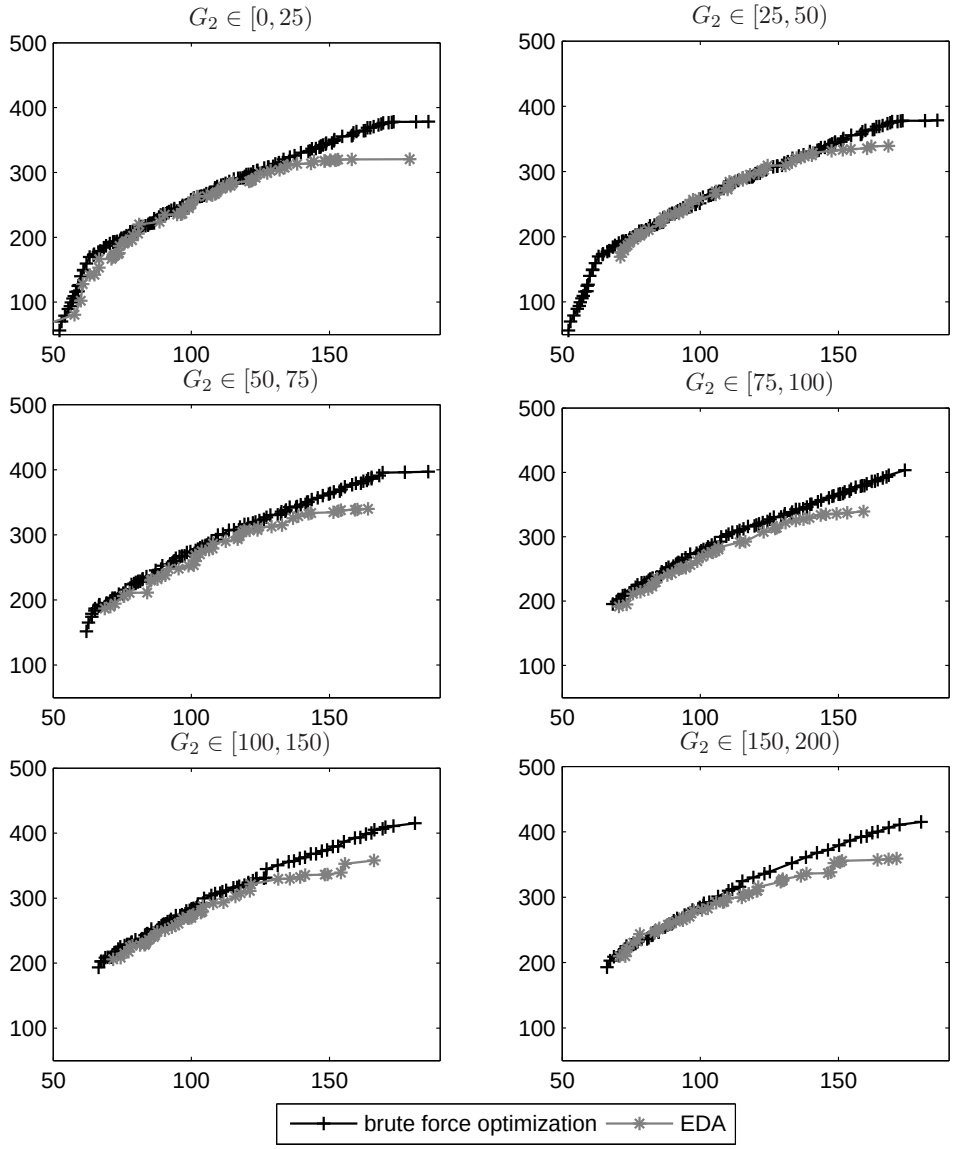


Figure 4.11: Pareto fronts obtained from multiple optimization runs including crossvalidation results and benchmark allocations from CHE case study

sulting throughput of the different patient groups, respectively. A boxplot graphically depicts five summary statistics. In each box the central mark corresponds to the median, the edges of the box depict the 25% and 75% quantiles. The whiskers indicate extreme values in the sample not consid-

ered outliers[1], if applicable the latter are plotted individually by a cross. Figure 4.12 shows that the optimized allocations comprise the entire range

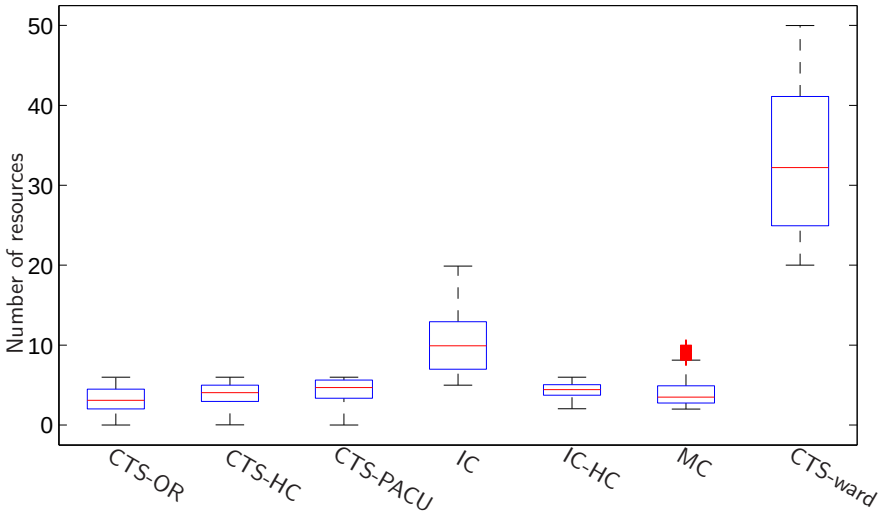


Figure 4.12: Boxplots of optimized allocation parameters

of possible allocation values, cf. Table 4.1. The different parts of the boxes, however, show the bulk of the allocation parameters ranges well within the resource bounds. As summarized in Table 4.2, only a small proportion the

Allocation parameters	**CTS-OR**	**CTS-HC**	**CTS-PACU**	**IC**	**IC-HC**	**MC**	**CTS-ward**
$r_u^* = r_u^{min}$	2.44%	1.17%	1.49%	5.66%	1.91%	12.67%	2.4%
$r_u^* = r_u^{max}$	6.12%	9.98%	25.51%	0.35%	9.91%	0.82%	1.95%

Table 4.2: Proportion (%) of optimal solutions $\mathbf{r}^*$ that take a parameter value r_u^* equal to the allocation bounds r_u^{min} and r_u^{max}, $u \in U$

optimized solutions feature parameter values that equal the upper and lower resource bounds, respectively. A concurrently minimal allocation at CTS-OR, CTS-HC and CTS-PACU occurs in only 0.3% of the allocations. No optimal allocation features an overall minimal allocation. The optimal allocations trend towards larger CTS-PACU units which can be attributed to the patients being relatively unsusceptible to complications requiring ICU care

[1]In a boxplot, points are considered as outliers if they lie outside 1.5 times the interquartile range, defined as the difference between the 25% and 75% quantile.

and that the CTS-PACU is dedicated to type II patients, cf. Section 2.3.5.

Moreover, the whiskers in Figure 4.12 indicate a right-skewed tendency of the IC, MC and CTS-ward allocation parameters, which means that the optimal solutions typically feature "small" allocations at these care units. The distributions of CTS-HC and CTS-PACU parameters are rather left-skewed, whereas the CTS-OR and IC-HC parameters appear to be symmetrically distributed. Also, the small range of IC-HC allocation parameter values is notable. These results indicate that considerable performance improvements by changing and coordinating the resource allocations at the different units can be achieved without large additional investments in expensive capacity, especially at the ICU.

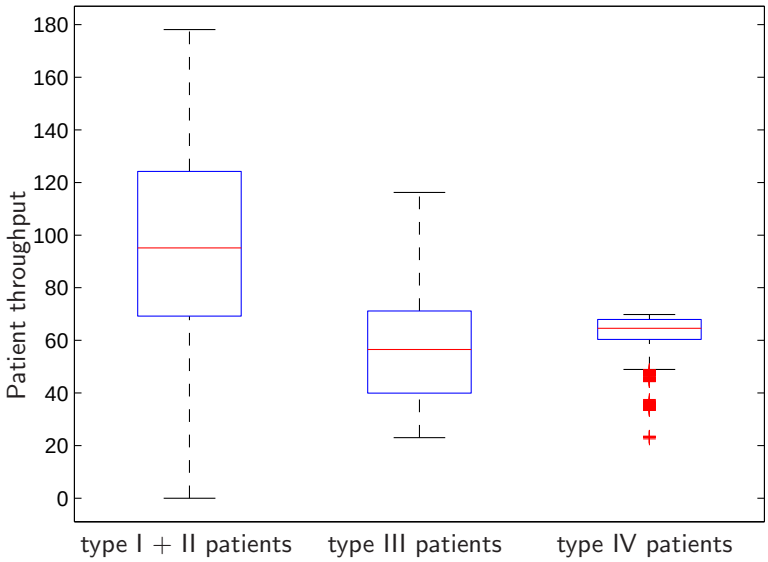


Figure 4.13: Boxplots of patient throughput resulting from optimized resource allocations

Figure 4.13 depicts the patient throughput resulting from the optimized resource allocations for the different patient groups. The current CHE resource allocation provides for on average 140, 40 and 70 CTS patients (type I+II), type III and type IV patients, respectively, cf. Table 2.5. Compared to this patient mix, the optimized solutions obtained by our approach still feature heterogeneous patient mix as the type III and IV patient throughput are positive and an omission of type I+II patient flow occurs in less than 1% of the allocations. In general, the resulting patient flows tend towards a

slightly reduced patient flow of type I + II patients, an increased proportion of type III patients and comparable type IV patient throughput. The bulk of optimized solutions feature positive resource allocations at CTS-PACU and CTS-HC provide for a heterogeneous throughput of type I+II patients as these resource allocations play a decisive roll in the corresponding patient admissions, cf. Section 2.4.

A patient throughput that is within 10% of the patient throughput for type I+II, III and IV patients given the current CHE allocation is achieved by 15.85%, 12.41% and 63.83% of the allocations, respectively. Moreover, about 2% of the allocations feature an overall patient mix that differs no more than 10% from the basic mix.

We can observe that similarly to the CHE case study, the proportion of patient types in the achieved total patient throughput differs noticeably, with the greatest throughput being achieved for type I+II patients. Also, a large range in patient throughput, especially for type I+II patients, can be noted which can be attributed to the stochastics of the corresponding patient pathways and the complex interactions between patient flows. Patient pathways of type I and II are especially susceptible to the overall resource allocation as they involve almost all units in the system among which multiple units are shared among the different patient types. For type III patients the stochastic arrival process affect the throughput, but also the large range of the resource allocations plays an important role as the resource availability is taken into account for admission decisions, cf. Section 2.3.

Intermediate conclusions We can conclude that the number of resources allocated to the different units mostly do not need to be maxed out in the optimal allocations. Rather the coordination of resource allocations considered our approach at the different units provides for optimal performance. Moreover, the proposed MO optimization approach provides for a heterogeneous patient mix that differs somewhat from the patient mix at the case study hospital which can be attributed to the dependency of patient flows on the resource allocation in the system.

4.6 Conclusions

In this chapter we addressed the multi-objective optimization for hospital resource management using evolutionary algorithms. In our model we considered the simultaneous allocation of multiple types of hospital beds and OR time slots in a network of care units. Due to the complexity of the

application domain, we used a state-of-the-art evolutionary MO technique, SDR-AVS-𝕄IDℍA. The fitness of the solutions was determined using the large-scale simulation described in Chapter 2. To the best of our knowledge this is the first approach in hospital resource management that combines MO optimization using an evolutionary algorithm with a realistic and validated simulation model that considers multiple types of hospital resources. In our experiments we analyzed the convergence behavior of SDR-AVS-𝕄IDℍA for different population sizes and determined the minimally required runtime which allows for the otherwise computationally too expensive optimization to be performed on a single PC. Moreover, our results showed that the benchmark allocations obtained from the CHE case study could be considerably improved using the optimized allocations. This improvement is possible through the concurrent optimization of resource allocations at the multiple hospital units which provides for a match between demand for care and available capacity on the care network level. Furthermore, the resource allocations obtained by our approach feature a heterogeneous patient mix and the majority of optimized solutions does not require large investments for additional capacity.

Our results showed that the optimized resource allocations provide for a heterogeneous patient mix. As the patient flows strongly depend on the resource allocation, the different trade-off optimal solutions result in considerably different patient mixes. For controlling the patient throughput, the decision maker should therefore take this criterion into consideration when choosing an optimized resource allocation to be implemented in practice. Several optimized allocations found by the EDA both feature a patient mix that is comparable to the CHE mix among and Pareto dominate the current CHE allocation.

The proposed approach is very flexible as the model parameters can be easily adjusted to different pathway and hospital settings. We demonstrated that a complex and realistic simulation in combination with state-of-the-art EDA can make an important contribution and achieve an improvement for complex real-world MO problems as in hospital patient flow logistics. Moreover, optimized allocation parameters can be obtained within feasible computation time and the results yield very good approximations to the globally Pareto optimal solutions.

In this chapter we tackle the multi-objective stochastic optimization problem of hospital resource management by optimizing the mean value of each of the three objective functions. This is a commonly applied approach for decision-making under uncertainty. As the underlying probability distribution of the different objective functions is unknown a priori, we used

a complex simulation to evaluate the performance of the different resource allocations. In this problem setting numerous factors contribute to the underlying distributions of the resulting outcome measures. This supports the assumption that total patient throughput, resource costs and back-up capacity usage follow a normal distribution. Considering the mean value of the underlying distributions in the optimization then provides for the resource allocation decisions being optimized for the bulk of the cases.

The evolutionary multi-objective optimization approach presented in this chapter can do well in a relatively stable environment. This condition, however, may not generally hold in hospitals. Due to the stochastic patient arrivals and pathways, as well as the disruptive nature of patient transfer scheduling, cf. Chapter 1 and Chapter 2, the resource usage behaves like a stochastic process, cf. Chapter 3. Therefore, the approach presented in this chapter will be extended in Chapter 5 to enable the adaptive allocation of resource to switch and track changes in the environment dynamically.

Chapter 5

Policy optimization for adaptive hospital resource management

In this chapter, we extend the evolutionary multi-objective optimization for hospital resource management presented in Chapter 4 to facilitate adaptive resource allocations. We propose a policy optimization approach where the resource allocation is determined using policies, i.e. parameterized functions that return an allocation decision given the current situation. The parameters of the different allocation policies are optimized using a multi-objective evolutionary algorithm (MOEA). Moreover, we present a way of performing anticipation in online dynamic multiobjective optimization using allocation policies to tackle the problem of time-dependence, i.e. decisions taken now have consequences in the future. The policies designed in this chapter enable the adaptive allocation of resources and the offline optimization of the policy parameters. Moreover, the policies are designed to make the solutions understandable to hospital professionals which is important for implementing the adaptive policies in practice. We demonstrate that these techniques can be applied to a real-world problem setting and show that the results outperform the optimized resource allocations obtained in Chapter 4. The use of anticipation in the allocation policies is found to lead to substantial improvements. A preliminary version of Chapter 4 and this chapter appeared as [48]. A publication on the basis of the anticipation approach presented in this chapter with contributions of Chapter 3 will appear as [49].

5.1 Introduction

In the previous chapter we presented an evolutionary multi-objective optimization for hospital resource management. The allocations considered in Chapter 4 allocate a fixed number of resources to the different hospital units which is typically employed by hospitals and is also current practice at the CHE case study hospital. Such a static allocation of resources is particularly suitable for hospital environments that are relatively stable. This condition, however, may not hold in clinic settings where the demand for care fluctuates over time. Often, stochastic patient arrivals and pathways as well as the highly disruptive scheduling of patient transfers cause the resource usage to behave like a stochastic process as the analysis in Chapter 3 has shown. Moreover, multiple stochastic patient pathways have to be taken into consideration that often share resources. Thus, adjusting the resource capacity to match the changing demand for care is a highly complex and dynamic problem.

Due to the dynamics of the problem, the optimization typically needs to be performed online, i.e. as time goes by. The difficulty involved in practical online dynamic optimization is the inherent time-dependence. This means that an allocation decision taken now has consequences on the future. Solving this problem myopically, i.e. considering only the current situation, can lead to inferior results over time. Consider, for example, the removal of an ICU bed due to low utilization of the allocated beds. Removing an ICU bed may cause a bottleneck at a postoperative care unit where patients have to remain longer than planned which in turn may cause cancelations of future surgeries. Moreover, multiple conflicting objectives need to be taken into consideration at the same time, cf. Chapter 4. Thus, the multiple objectives and the time-dependence are important to be taken into account which further complicate the optimization.

For adaptive hospital resource management, we present an optimization approach where the resource allocation is determined using policies, i.e. parameterized functions that return an allocation decision given the current situation. The policies' parameters are optimized using a multi-objective (MO) evolutionary algorithm (MOEA). MOEAs have been shown to be very powerful for MO optimization problems [14, 20, 32]. Moreover, MOEAs are an efficient approximation technique for complex real-world problems where the objectives are not clear mathematical functions, but rather a complex simulation, as in the situation considered in this thesis. The advantage of using policies to solve stochastic dynamic optimization problems is that only one strategy has to be optimized that can be applied to a set of scenarios

in the simulation. In cooperation with domain experts from the case study hospital, cf. Section 1.3.1, we designed policies that enable adaptive resource allocations that are implementable in hospital practice. Therefore, the policies can be easily understood by health care professionals which is important for the understanding in practice and real-life implementation.

Thus, adaptive hospital resource management is a complex and dynamic problem that requires state-of-the-art techniques from dynamic MO research. Specifically, we introduce a combination of policy optimization and the evolutionary multi-objective approach presented in Chapter 4 using the SDR-AVS-$\mathbb{MIDEA}$ algorithm [15]. We demonstrate the applicability of the our policy optimization approach using the CHE simulation instance, cf. Section 2.3.5, and show that the more dynamic resource allocations can further improve the optimized resource allocations in Chapter 4. To better tackle time-dependence, we also present an approach to anticipation in online dynamic MO optimization using allocation policies. For this purpose we apply the forward simulation prediction approach presented in Chapter 3 and show that better results can be obtained taking the future resource occupancy into account.

The remainder of this chapter is organized as follows. First, we briefly discuss related previous work in Section 5.2. Then, we extend the model for hospital resource management presented in Section 4.3 to adaptive resource allocations in Section 5.3. Next, the adaptive allocation policies are presented in Section 5.4. The experiments are reported in Section 5.5. We end this chapter with our conclusions.

5.2 Related work

Dynamic MO optimization has been addressed in few earlier studies in the field of evolutionary algorithms, especially considering stochastic environments. Moreover, the current literature consists mostly of first definitions and algorithms. The approach presented in [32] is developed for seldom random changes of the environment and requires optimization from scratch if a change in the environment is detected. However, this approach is not suitable to be applied in the problem setting considered in this thesis because the stochastic and complex patient pathways provide for frequent changes in resource occupancy and thus resource requirements. Our approach uses policy optimization and therefore does not need to be re-optimized for each situation. Moreover, it can handle also frequent changes of the environment because the strategies describe what to do in any situation. In [20]

the performance of the Non-dominated Sorting Genetic Algorithm version 2 (NSGA2) is evaluated in a stochastic setting for artificial objective functions. The changes of the objective functions, however, are unrelated to the choices made for the problem variables. In our work, we use objective functions for a real-world application where time-dependence is an important source of dynamism and the previous approaches are therefore not applicable.

Currently, there is no literature on performing anticipation in the optimization of multiple dynamically-changing objectives yet.

Related work on hospital resource management in the operations research and operations management literature is equivalent to the earlier work discussed in Chapter 4. However, none of the approaches presented in the literature has yet considered adaptive resource allocations which is the focus of this chapter.

5.3 Model

As argued above, the adaptive hospital resource management problem is a dynamic multi-objective optimization problem. In this section, we describe the key concepts of dynamic multi-objective optimization and the policy optimization approach taken in this chapter. For an outline of multi-objective optimization and the relevant key-concepts the reader is referred to Section 4.3.3.

5.3.1 Dynamic multi-objective optimization

As described in the domain and patient flow model in Chapter 1, Section 1.1.3, we consider discrete equidistant decision moments denoted by $t_i \in T$ with $t_{i-1} < t_i$ for $i = 1, \ldots, n-1$. Typically, t_i would be in steps of days for adaptive resource management. Furthermore, we denote the prediction horizon of future resource usage by $h \in \mathbb{N}_0$ with $0 \leq h \leq n-1$, cf. Chapter 3.

Similarly to Chapter 4, we consider the number of allocated resources as free decision variables in the optimization problem at hand. Let $\mathbf{r}(t_i) = (r_u(t_i),\, u \in U)$ denote the number of resource allocated to unit u at time $t_i \in T$.

Similarly to the model for fixed resource management presented in Section 4.3, the dynamic optimization of the resource allocation during period

T can be formalized mathematically as

$$\min_{\mathbf{r}(t_k)} \left\{ \int_{t^0}^{t^{n-1}} G_{t_k}(\mathbf{r}(t_k))\, dt_k \right\} \tag{5.1}$$

subject to

$$\forall u \in U \, \forall t_i \in T : \, r_u(t_i) \in \mathbb{N} \cap [r_u^{min}, r_u^{max}], \tag{5.2}$$

where the function G_{t_k} is multi-objective,

$$G_{t_k}(\mathbf{r}(t_k)) = (-G_{0,t_k}(\mathbf{r}(t_k)), G_{1,t_k}(\mathbf{r}(t_k)), G_{2,t_k}(\mathbf{r}(t_k))).$$

G_{0,t_k}, G_{1,t_k} and G_{2,t_k} refer to the mean total patient throughput, resource costs and back-up capacity usage up to time t_k that result from simulating $\mathbf{r}(t_k)$, $k = 0, \ldots, n-1$ in the agent-based simulation, cf. Chapter 2 and Section 4.3.2. The integral in (5.1) represents the optimization over time. Here, we take the integral of a multi-objective function to be the multi-objective function of the integrals, i.e. (5.1) is defined as

$$\min_{\mathbf{r}(t_k)} \left\{ \int_{t^0}^{t^{n-1}} -G_0(\mathbf{r}(t_k))\, dt_k, \int_{t^0}^{t^{n-1}} G_1(\mathbf{r}(t_k))\, dt_k, \int_{t^0}^{t^{n-1}} G_2(\mathbf{r}(t_k))\, dt_k \right\}. \tag{5.3}$$

Solving this problem online means that at any point in time t_i, the objectives G_{j,t_i} cannot be evaluated for any t_k beyond the current time t_i, i.e. $t_k > t_i$. Myopically minimizing (5.3) amounts to solving the following multi-objective problem repeatedly for $t_i \in T$

$$\min_{\mathbf{r}(t_i)} \left\{ -G_{0,t_i}(\mathbf{r}(t_i)), G_{1,t_i}(\mathbf{r}(t_i)), G_{2,t_i}(\mathbf{r}(t_i)) \right\}. \tag{5.4}$$

Following this myopic approach two important problems arise. First, in a real-world setting, the variables to optimize are decision variables. This means that at any point in time only one solutions, i.e. one point on the corresponding Pareto front corresponding to a resource allocation, can be selected as the decision to be taken. How to select solutions over time is still an open issue in dynamic multi-objective optimization.

Second, decisions may have future consequences. This means that a value of the optimal solution at time t_k may depend on previous decisions $\mathbf{r}(t_j)$, $t_j < t_k$. If only the current situation is taken into account, the decision that leads to a Pareto optimal solution for the current time period is optimal.

Considering the entire time span T, however, an optimal solution may be suboptimal from a myopic point of view.

To tackle these problems we present a policy optimization approach in combination with a MOEA as advocated in [13] which will be explained below.

5.3.2 Policy optimization approach

Policies are parameterized functions that return a decision for any given situation, denoted here by s. Specifically, we use allocation policies, $\pi(t_i, s) = (\pi_u(t_i, s), u \in U)$, that determine the number of resources, $r_u(t_i)$, allocated to hospital unit $u \in U$ at time $t_i \in T$ given the current situation situation s such that

$$r_u(t_i) = \pi_u(t_i, s) \; \forall u \in U, \forall t_i \in T. \tag{5.5}$$

The focus of the optimization is on finding the best combination of policy parameters within a fixed policy equation structure to minimize the objective functions described in Section 5.3.1.

Adaptive policies

The simulation is the run with the policy and whenever an allocation decision is required, the policy is evaluated that returns the decision to be taken. The policy defined above is adaptive as the outcome depends on the current situation s and thus automatically adapts to s. For example, for a hospital care unit with currently $r_u(t_i)$ allocated beds, an adaptive allocation policy can state that if the utilization is below a value $util_{min}$, a bed should be removed from the unit to reduce the resource costs. This allocation policy thus adapts to the current utilization during simulation.

Offline optimization of online allocation decisions

The proposed adaptive policy optimization approach allows the resources (i.e. decision variables) to switch and track changes in the environment (i.e. the optimization problem) dynamically. The advantage of this approach for solving stochastic dynamic optimization problems is that only one policy has to be optimized that can be applied to multiple simulation scenarios or runs instead of determining an optimal allocation decision at every decision moment $t_i \in T$. Also, our approach does not require further optimization during the simulation run. Therefore, the policy optimization approach allows us to perform the multi-objective optimization in an offline fashion

using the agent-based simulation described in Chapter 2, cf. Figure 2.1. Thus, MO techniques can be applied in a straightforward fashion to solve this dynamic problem which yields an additional advantage. Moreover, if the policies are designed properly, the policies can be easily understood by domain experts which is important for the final implementation in practice.

Anticipation in policy optimization

As we argued above, resource allocation decisions taken now may have future consequences. This means that an optimal allocation decision now may depend on the previous decisions $r_u(t_k)$, $t_j < t_k$. In order to take the future into account, prediction information needs to be considered in the decision making. Here, prediction information can be a statistical model that is learned from past observations, cf. Section 3.5 or simulation if available. In this chapter, we will use the latter prediction approach due to its higher accuracy and apply forward simulation as discussed in Section 3.4.

In order to obtain anticipation in offline policy optimization, the policies need to be designed to be anticipatory. This means that inside the policy, the time interval $[t_i, t_{i+h}]$ needs to be taken into account. Our approach is to consider prediction information over $[t_i, t_{i+h}]$ in the policy which provides for the returned allocation decision to be determined on the basis of the values predicted in this time interval. An adaptive policy that incorporates prediction information is in the remainder referred to as anticipatory policy.

Consider the allocation policy example outlined above. Prediction information may indicate that within $[t_i, t_{i+h}]$ multiple patient transfers from other care units are to be expected which provide for a high resource occupancy and thus utilization. Using this prediction information in the policy then provides to refrain from the capacity reduction as would otherwise be done due to the current low utilization.

When evaluating an anticipatory policy at the current time t_i, forward simulation is used to predict the resource occupancy during the time interval $[t_i, t_{i+h}]$. During $[t_i, t_{i+h}]$ an adaptive policy needs to be used to make allocation decisions. Using an anticipatory policy involves a possibly indefinite recursive calling between forward simulation and the policy, depending on the prediction horizon h. If h is large, then a sufficiently good approximation of the future could be achieved through applying the anticipatory policy during $[t_i, t_{i+h'}]$ with $t_{i+h'} < t_{i+h}$ and a policy that does not incorporate prediction information during $[t_{i+h'}, t_{i+h}]$ provided that the influence of time-dependence is not too strong.

5.4 Adaptive policies for hospital resource management

The adaptive policies described below were developed in cooperation with domain experts from the CHE case study hospital, cf. Chapter 1, Section 1.3.1. Therefore, the policies can be easily understood by health care professionals making the practical implementation of the optimized solutions much easier.

In the following, the adaptive allocation policies studied in this chapter are presented. Moreover, a mechanism for exchanging resources among hospital care units is described that enables the implementation of an adaptive resource allocation in practice.

5.4.1 Adaptive state-dependent allocation policies

As argued above, we propose adaptive allocation policies that return an allocation decision for the units in the network, given the problem variables or state. For determining the state we consider two cases:

non-anticipatory policy: the state is determined by the *current* occupancy information available at the decision moment,

anticipatory policy: the state incorporates information on the *predicted* resource usage during $[t_i, t_{i+h}]$ given the current resource occupancy.

Below, the state representation, the policy and its usage for dynamic resource allocation are described.

State description

The state description of the two cases outlined above are given below.

Current resource occupancy information In this situation, the state at unit u, $s_u^{now}(t_i)$, is based on the occupancy information available at decision moment t_i and is determined by the resource utilization rate at u, i.e. the ratio between the occupied capacity[1] at the start of day t_i and the

[1]Note that due to the possible usage of back-up capacity the (predicted) occupied capacity may exceed the allocated resources, thus the (predicted) utilization rate may be greater than 1 for some units $u \in U$.

resource capacity, $r_u(t_i^-)$, just before the adjustment at t_i, denoted by t_i^-. Formally, we have $s_u^{now} : T' \to \mathbb{R}_0^+$, $u \in U$, with

$$s_u^{now}(t_i) = \frac{\#\text{occupied resources at unit } u \text{ at start of day } t_i}{r_u(t_i^-)}. \tag{5.6}$$

For some postoperative care units, the state at the start of day t_i defined by (5.6) may not be representative for the resource occupancy during the remainder of day t_i. Consider for example the CHE case study. At the CTS-PACU, the beds are available only for a couple of hours during the day and are opened just before the first surgeries are expected to be completed, cf. Section 2.3.5, Table 2.1 on page 45. At the CTS-HC not all beds may be occupied, e.g. if a type I surgery has been canceled on day t_{i-1} due to a type IV admission but the type IV patient has been transferred to another unit prior to t_i^-. However, according to the OR scheme all beds will be occupied during day t_i. Instead of determining the state by (5.6), we propose to use a flow-based heuristic to determine the state for these units. The heuristic determines the utilized capacity for day t_i as the number of occupied resources at time t_i^- minus the expected patient outflow plus the expected inflow (determined by the OR scheme) for day t_i divided by the current resource allocation. In the CHE case study, the utilized capacity for the CTS-PACU can be further simplified such that the utilization is determined by the fraction of the OR scheme for type II patients and the current resource allocation.

Predicted resource occupancy information The state at unit u determined using prediction over the time period $[t_i, t_{i+h}]$, $s_u^{now+h}(t_i)$ with $h \in \mathbb{N}_0$, is determined by the predicted mean resource utilization rate[1] at u during the period $[t_i, t_{i+h}]$, i.e. the ratio between the mean predicted number of occupied resources averaged over the period $[t_i, t_{i+h}]$ and the resource capacity, $r_u(t_i^-)$, just before the adjustment at t_i. The average predicted number of occupied resources is determined by the predicted density function, $\hat{f}^u_{t_k;\mathbf{a}_{[t_i,t_{i+h}]}}$ derived from the predicted empirical cumulative probability distribution function $\hat{F}^u_{t_k;\mathbf{a}_{[t_i,t_{i+h}]}}$, $k = i, \ldots, i+h$, cf. Section 3.3. Then, s_u^{now+h} is determined by

$$s_u^{now+h}(t_i) = \frac{\frac{1}{h+1}\sum_{k=i}^{i+h}\int_0^\infty y\hat{f}^u_{t_k;\mathbf{a}_{[t_i,t_{i+h}]}}(y)dy}{r_u(t_i^-)} \tag{5.7}$$

Due to the temporary resource availability at the CTS-PACU, the corresponding resource occupancy distribution is a two-heaped distribution.

Thus, the mean is not representative for the resource usage during the period $[t_i, t_{i+h}]$. Therefore, the predicted state for the CTS-PACU is determined by the flow-based heuristic described in the previous paragraph.

State-dependent allocation policy

The proposed state-dependent policy for the resource allocation problem at hand is designed for the state being determined both by the current as well as the predicted occupancy information. In the latter case, we refer to the state-dependent policy as anticipatory policy. Below, the generic state information used at time t_i for unit u is therefore denoted by $\cdot$ in $s_u^{\cdot}(t_i)$, $t_i \in T$, $u \in U$.

A state-dependent allocation policy, denoted by $(\pi_u(t_i, s_u^{\cdot}), t_i \in T, u \in U)$, is determined by five parameters: a base resource allocation, r_u^{base}, two adjustments, r_u^{decr} and r_u^{incr}, and two utilization thresholds, $\mathcal{UT}_u^{decr}$,$\mathcal{UT}_u^{incr}$ with $\mathcal{UT}_u^{decr} \leq \mathcal{UT}_u^{incr}$. We use an iterative step-function $\pi : T \times \mathbb{R}_0^+ \to \mathbb{N}^{|U|}$ given as

$$\pi_u(t_i, s_u^{\cdot}) = \begin{cases} \max\{r_u^{min}, r_u(t_i^-) - r_u^{decr}\} & \text{, if } s_u^{\cdot}(t_i) < \mathcal{UT}_u^{decr} \\ r_u(t_i^-) & \text{, if } s_u^{\cdot}(t_i) \in [\mathcal{UT}_u^{decr}, \mathcal{UT}_u^{incr}] \\ \min\{r_u^{max}, r_u(t_i^-) + r_u^{incr}\} & \text{, otherwise} \end{cases} \tag{5.8}$$

for $t_1, \ldots, t_{n-1}$ and

$$\pi_u(t_0, s_u^{\cdot}) = r_u^{base}, \tag{5.9}$$

with $\pi_u(t_i, s_u^{\cdot}) \in [r_u^{min}, r_u^{max}]$ $\forall t_i \in T, u \in U$. In (5.8) the current resources allocation, $r_u(t_i^-)$, is decreased by r_u^{decr} if the resource utilization rate is below the threshold $\mathcal{UT}_u^{decr}$. If the utilization rate is above $\mathcal{UT}_u^{incr}$, $r_u(t_i^-)$ is increased by r_u^{incr}. Otherwise, the current allocation remains unchanged. Note that the policy specifies the allocation at the different units independently.

For the policy-based allocation approach proposed in this chapter, the constraint of the optimization problem specified by (5.2) can thus be replaced by

$$\forall u \in U : r_u^{base} \in \mathbb{N} \cap [r_u^{min}, r_u^{max}], \tag{5.10}$$

$$\forall u \in U \, \forall t_i \in T' : s_u^{\cdot}(t_i) \in \mathbb{R}_0^+, \tag{5.11}$$

$$\forall u \in U : r_u^{decr}, r_u^{incr} \in [0, 5] \tag{5.12}$$

$$\forall u \in U : \mathcal{UT}_u^{decr} \in [0, 1], \mathcal{UT}_u^{incr} \in [\mathcal{UT}_u^{decr}, \mathcal{UT}_u^{decr} + 1]. \tag{5.13}$$

As large adjustments are not desirable for hospital management, a maximal adjustment of 5 beds was chosen. Based on preliminary runs, an upper bound of 2 for $\mathcal{UT}_u^{incr}$ appeared to be more than sufficient.

5.4.2 Bed exchange mechanism

In the adaptive allocation policy described in Section 5.4.1, a large supply and stock of beds is assumed which enables the concurrent in- and decrease in resource capacity at the different units. In reality, however, bed availability is restricted by the available staff, in particular the number of personnel needed per bed at a specific unit. Staff schedules need to be fixed at least several weeks in advance. The use of stand-by personnel is not common in the hospital domain. Therefore, a direct implementation of the policy described in Section 5.4.1 is often not practically feasible. To enable adaptive resource allocation in hospitals, we propose an exchange mechanism that is based on fixed personnel resources. The resources are exchanged among the hospital units to meet the current local need at the different units.

Here, $\pi_u(t_i, s_u^{\cdot})$ denotes the number of resources *required* by unit u at time t_i, determined by (5.8) based on the state $s_u^{\cdot}(t_i)$. The fixed personnel resources are determined by (5.9), i.e. the base allocation $r_u^{base}, u \in U$. The resource allocation at time t_i, $r_u(t)$, is set by the mechanism below and not by (5.8).

For exchanging beds we distinguish different care levels based on the intensity of care and monitoring provided at a hospital care unit, cf. Chapter 1, Section 1.1.1. Here, the classification of hospital units into L care levels is focussed on the staffing requirements of the resources and the mandatory level of training of the personnel such that the corresponding staffed resources are comparable and interchangeable within a care level. Therefore, the care level of a unit is also linked to the unit costs of the resources, i.e. a high care level of a care unit corresponds to high unit resources costs. We denote the set of hospital care units classified with care level l by U_l with $l = 1, \ldots, L$ where the staffing requirements and training level of the personnel decreases for increasing value of l.

From the application domain four rules arise for feasible bed exchanges:

R1 Due to staff training and physical requirements (e.g. access availability to the isolated electric power system in the hospital), beds can only be exchanged within the same or between adjacent care levels.

R2 Due to the staff assigned to a bed, shifting one bed from level l to level $l + 1$ yields ρ_l beds at $l + 1$ for $l = 1, \ldots, L - 1$.

R3 Due to the personnel required to operate a bed, only a multiple of ρ_{l-1} beds can be shifted from level l to level $l-1$ for $l = 2, \ldots, L$ (i.e. the reverse of **R2**).

R4 In order to prevent upgrading of resources to a higher care level, an upper bound is imposed on the number of resources that can be shifted from level l to level $l-1$ which is implied by the initial allocation r_u^{base}, $u \in U$. The bounds are given by

$$\sum_{u \in U_1} (r_u^{base} - r_u(t_i^-)) \tag{5.14}$$

and

$$\sum_{u \in U_l} (r_u^{base} - r_u(t_i^-)) + \rho_{l-1} \cdot \sum_{u \in U_{l-1}} (r_u^{base} - r_u(t_i^-)) \tag{5.15}$$

for level 1 and $l = 2, \ldots, L$, respectively. Applying this rule implies that the possible resource exchanges are limited by the initial resource allocation taking the current allocation into account. For levels $l > 1$ also the adjacent higher care level is taken into account as higher level beds may be allocated at the lower care level according to rule **R2**.

Thus, rules **R1** and **R4** determine the general conditions for adjusting the resource allocation while rules **R2** and **R3** regulate the 'exchange rate' for shifting beds between different care levels.

A high-level outline of the mechanism is provided in Algorithm 3. The input of the bed exchange mechanism comprises the different units and their care levels, the resource need at the different units given by (5.8). Based on the required number of resources, determined by (5.8), the number of excess resources, $\mathcal{E}_l$, in care level l at time t_i is determined by

$$\mathcal{E}_l = \sum_{u \in U_l} \max\{0, r_u(t_i^-) - \pi_u(t_i, s_u^{\cdot})\},\ l = 1, \ldots, L. \tag{5.16}$$

Similarly and taking rule **R4** into account, the required number of resources at level l is given by

$$\mathcal{R}_1 = \min\{(5.14), \sum_{u \in U_1} \max\{0, \pi_u(t_i, s_u^{\cdot}) - r_u(t_i^-)\}\}, \tag{5.17}$$

and

$$\mathcal{R}_l = \min\{(5.15), \sum_{u \in U_l} \max\{0, \pi_u(t_i, s_u^{\cdot}) - r_u(t_i^-)\}\},\ l = 2, \ldots, L. \tag{5.18}$$

Algorithm 3: Pseudo-code description of the bed exchange mechanism

Input: Sets of hospital units, U_l, of level $l = 1, \ldots, L$
$\pi_u(t_i, s_u)\, \forall u \in \bigcup_l U_l$, given by (5.8)
$\mathcal{E}_l$, $l = 1, \ldots, L$, determined by (5.16)
$\mathcal{R}_l$, $l = 1, \ldots, L$, determined by (5.17) and (5.18)
Result: $r_u(t_i)\, \forall u \in U$

foreach *care level* $l = 1, \ldots, L$ **do**
  if $\mathcal{E}_l > 0$ *and* $\mathcal{R}_l > 0$ **then**
    Shift $\min\{|\mathcal{E}_l|, |\mathcal{R}_l|\}$ beds to unit(s) of level l and decrease $\mathcal{E}_l$ and $\mathcal{R}_l$ accordingly;
  if $\mathcal{E}_l > 0$ *and* $\mathcal{R}_{l+1} > 0$ **then**
    Shift $\min\{|\mathcal{E}_l|, \lfloor|\mathcal{R}_{l+1}|/\rho_l\rfloor\}$ beds to unit(s) of level $l+1$ and decrease $\mathcal{E}_l$ and $\mathcal{R}_{l+1}$ accordingly (applying rule R1 + R2);
  if $\mathcal{R}_l > 0$ *and* $\mathcal{E}_{l+1} > 0$ **then**
    Shift $\min\{|\mathcal{R}_l|, \lfloor|\mathcal{E}_{l+1}|/\rho_l\rfloor\}$ beds to unit(s) of level l and decrease $\mathcal{R}_l$ and $\mathcal{E}_{l+1}$ accordingly (applying rule R1 + R3);
foreach *unit* $u \in \bigcup U_l$ *that was not yet considered* **do**
  Allocate $r_u(t_i^-)$ resources;

If the accumulated resource requirement exceeds the bounds given by (5.14) and (5.15), the corresponding number of required resources are decreased proportionately to $\pi_u(t_i, s_u^{\cdot})$.

According to the mechanism described in Algorithm 3 the different care levels are iteratively considered in ascending order. First, beds are shifted within level l. Then, level $l+1$ resources are shifted to level l if necessary. Subsequently, resources are exchanged between level $l+1$ and l. All exchanges are performed only if additional resources are required by a care unit of the considered care level, i.e. if $\pi_u(t_i, s_u^{\cdot}) > r_u(t_i^-)$, $u \in U_l$ and $l = 1, \ldots, L$. The care units are considered in ascending order of their care level l due to the decreasing resource costs. In general, the resource costs contribute at least partially to the fact that fewer resources with high associated costs are available in a hospital than resources with low resource costs. This is also the reason why back-up capacity for low cost resources is more readily available than for expensive hospital resources. Therefore, it is desirable for hospital management that units with a low value of l are considered first to exchange resources. Care units within a care level are selected in a random order for resource exchange in order to provide equal chances to the different units to adjust their capacity.

The actual shifting of resources is performed as follows. Two care units u and v with $\pi_u(t_i, s_u^{\cdot}) > r_u(t_i^-)$ and $\pi_v(t_i, s_v^{\cdot}) < r_v(t_i^-)$ are selected ran-

domly in the corresponding care level(s). If the number of resources u is willing to shift exceeds the number of resource required by v, then only the number of resources required by v is shifted and the remaining resources of u are considered for resource exchange with another unit. Depending on the specific rules and settings of the hospital where the exchange mechanism is to be applied, resources from a higher level may be split between units or combined from multiple units of the lower care level if rules **R2** or **R3** apply.

Through the mechanism, the implementation of (5.8) is extended with the above adjustments at time $t_i \in T$, depending on the interaction with other units. This complex interaction mechanism answers to reality, however, it further complicates the optimization of resource management. Therefore, a state-of-the-art technique is needed for this optimization.

5.5 Experiments and settings

In this section we describe the settings and the experiments that were performed to evaluate the proposed policy optimization approach. First, we provide the basic setup of SDR-AVS-𝕄IDEA and the agent-based simulation, cf. Chapter 2. Then, we determine the required subpopulation size and the number of evaluations to obtain high-quality optimization results at reduced computational costs for the policy optimization problem where current resource occupancy information is used. Then, we present the optimization results obtained for the non-anticipatory approach and analyze the optimized allocation policy parameters and the implications of their usage for hospital practice. Finally, we will analyze the results obtained for the anticipatory resource allocation policies, the optimized parameters and their resulting patient throughput.

5.5.1 Basic algorithmic setup

The MOEA we use is the SDR-AVS-𝕄IDEA [15]. The algorithm was shown to be an efficient optimization technique for MO optimization problems. A detailed description of SDR-AVS-𝕄IDEA is given in Section 4.4.3. The settings of the variation, selection, replacement parameters, the number of clusters, the maximally allowed number of generations and the discretization length used in the elitist archive in SDR-AVS-𝕄IDEA used in this chapter are consistent with the settings presented in Section 4.5.1 and will therefore be omitted here.

Since the set of globally Pareto-optimal solutions, P_S, for the adaptive resource management problem is unknown a priori and a brute force opti-

mization approach is infeasible due to the uncountable set of policy parameters, we have to approximate P_S. Similarly to Section 4.5, we approximate P_S by 10 independent runs of SDR-AVS-$\mathbb{MIDEA}$ with a large number of generations and a large population size, which is determined by (4.4) on page 128. Specifically, the maximal number of generations is set to 1600 and (4.4) results in a the subpopulation size of 130 given the 35 parameters to be optimized for the adaptive policies. Using P_S, we use the convergence measure, $D_{P_F \to \mathcal{S}}$, as defined in equation (4.3) on page127, to determine the minimally required subpopulation size, n^{subpop}, and number of evaluations in Section 5.5.3 and consider the corresponding Pareto fronts.

In the EDA representation, the individuals correspond to allocation policy parameters which are summarized in Table 5.1. Specifically, the genotype contains a parameter $\mathcal{T}_u \in [0,1]$ for $u \in U$ that is used to determine $\mathcal{UT}_u^{incr}$ by $\mathcal{UT}_u^{incr} = \mathcal{UT}_u^{decr} + \mathcal{T}_u$. This parametrization results to 35 real-valued parameters to be optimized using SDR-AVS-$\mathbb{MIDEA}$ for the different policy types. The bounds, r_u^{min} and r_u^{max}, for the resource allocations were obtained from domain experts from CHE and are given in Table 4.1 on page 127.

Parameters	**Description**	**Bounds**
r_u^{base}	The base allocation at unit $u \in U$	$[r_u^{min}, r_u^{max}]$
r_u^{decr}	The decremental adjustment at $u \in U$	$[0,5]$
r_u^{incr}	The incremental adjustment at $u \in U$	$[0,5]$
$\mathcal{UT}_u^{decr}$	The utilization threshold under which the current allocation is decreased by r_u^{decr}	$[0,1]$
$\mathcal{UT}_u^{incr}$	The utilization threshold above which the current allocation is increased by r_u^{incr}	$[0,2]$

Table 5.1: Summary of the adaptive resource allocation policy parameters for care units $u \in U$

5.5.2 Setup agent-based simulation

To evaluate the fitness of an adaptive policy, we run 10 simulation runs of 4 weeks after 12 weeks of warming-up. This setting results in a runtime of about 1.9 and 2.1 seconds per evaluation for the state-dependent policies and the exchange mechanism, respectively. For predicting the future resource occupancy we used 300 forward simulation scenarios as determined in Section 3.4.3 which results in a runtime of about 40 seconds per evaluation. In the simulation, the adaptive policies are applied at the start of

every adjustment period, which here is every day. Moveover, adjustments are performed only after the initial warming-up period. Based on preliminary results warming-up was found to be necessary in order to avoid early convergence to minimal allocations due to the empty hospital in the start of a simulation run.

For the evaluation of the exchange mechanism presented in Section 5.4.2, the simulation instance obtained from the CHE case study is used for setting the model parameters, cf. Chapter 2, Section 2.3.5. In accordance with domain experts from the CHE the classification of care units resulted in three care levels corresponding to the following care units: level 1 is the intensive care (IC), level 2 comprises the IC-HC, MC and the CTS-HC. The CTS-PACU unit is not considered for exchanging beds, since the beds are only staffed for a limited period of time and could thus not be allocated to a unit whose beds are open 24/7. Level 3 is the CTS-ward. The exchange rate between levels is determined by the ratio between the relative unit resource costs in the CHE case study, cf. Table 2.4 on page 54. Here, one level 1 bed equals two level 2 beds and one level 2 bed in turn equals two level 3 beds. In order to assess the effects of dynamically changing the resource allocation on the simulation performance, we consider the case of unconstrained admission control, cf. Chapter 3, Section 3.3.

For each policy, the simulation uses the same 10 random seeds to allow for a fair comparison. Similarly to the evaluation in Chapter 4, we perform crossvalidation of the optimized solutions by evaluating the solutions using 50 independent simulation runs to assess a possible overfit of the solutions for the fixed random seeds.

5.5.3 Setting the subpopulation size and the required number of evaluations

The guideline (4.4) on page 128 provides for a subpopulation size $n^{subpop} = 130$. Running SDR-AVS-MIDEA with $n^{subpop} = 130$ for 1600 generations results in a runtime of approximately 30 hours for adaptive policy optimization[2] on a high-performance computer cluster using 40 nodes running at speeds between 1.4Ghz and 2.2Ghz. Similarly to Chapter 4, the largest part of this runtime is used for evaluating the fitness of the adaptive policies using the simulation. Taking the considerations concerning the sine-objective guideline (4.4) and diversity-preserving selection in SDR-AVS-MIDEA, cf. Section 4.5.3, we varied the subpopulation size. Moreover, the

[2]Not including prediction in the state calculation.

currently achieved runtime is not feasible if the proposed approach is to be applied in a hospital setting where the optimization has to be performed regularly on a single PC. Therefore, we determine the number of computational resources that is required in order to obtain solutions that are reasonably close to the Pareto-optimal solutions and show sufficient diversity after a reasonable runtime. In order to a broad spectrum of subpopulation sizes, we varied n^{subpop} between 30, 80 and 130.

Similarly to the convergence results presented in Section 4.5.4, the corresponding Pareto-fronts are depicted with G_1 and G_0 values plotted on the horizontal and vertical axes, respectively, for predefined intervals of G_2 values with $G_2 \in [0, 25]$, $[25, 50]$, etc. Then, we discuss the results obtained form the different policies proposed in Section 5.4.

Determining the required subpopulation size

The initial size of the subpopulations was set to 130 as determined by (4.4). For our evaluations this population size was varied between 30, 80 and 130 for the adaptive policies and the exchange mechanism, respectively. These values were chosen such that a broad spectrum of population sizes is evaluated similar to the scale used in Section 4.5.4. The convergence graphs corresponding to the different subpopulation sizes are shown in Figure 5.1 and Figure 5.2 for the state-dependent policies and the bed exchange mechanism, respectively. For all policies and subpopulation sizes a steep decline of $D_{P_F \to \mathcal{S}}$ is to be noted in the first 100 generations, after this the decrease in $D_{P_F \to \mathcal{S}}$ is reduced. In the first 100 generations about 90% of the convergence of SDR-AVS-$\mathbb{MIDEA}$ is achieved. Moreover, the subpopulation sizes determined by (4.4) achieve a better and faster convergence compared to the smaller population sizes for all policy types.

For the state-dependent policies, a considerable difference in convergence can be noted between the different subpopulation sizes. The final $D_{P_F \to \mathcal{S}}$ value after 1600 generations for $n^{subpop} = 30$ and 80 amounts to about 6.4 and 3.1, respectively. Compared to the convergence results in Section 4.5.4, the policy optimization with $n^{subpop} = 130$ converges slightly faster with a $D_{P_F \to \mathcal{S}}$ value of less than 1 being achieved after about 800 generations which for the allocations in Chapter 4 takes about 100 generations longer for with $n^{subpop} = 50$. Also, SDR-AVS-$\mathbb{MIDEA}$ converges faster for subpopulation sizes $n^{subpop} = 30$ and 80 than for the optimization in Chapter 4. Moreover, we observed from the data of the obtained Pareto optimal state-dependent policies that the number of non-dominated solutions retained in the elitist archive obtained for $n^{subpop} = 130$ decreases to about 65% and 84% for

$n^{subpop} = 30$ and 80, respectively.

While the exchange mechanism requires the same number of parameters to be optimized as the state-dependent allocation policies, the additional interaction between the units to determine the resource allocation slows down the convergence, especially for the smaller subpopulation sizes. Also, it makes finding good solutions harder. The final value of $D_{P_F \to \mathcal{S}}$ remains at a high level for subpopulation sizes of 30 and 80. For $n^{subpop} = 30$ the final convergence performance is slightly worse compared to the state-dependent policy optimization, whereas for $n^{subpop} = 80$ the final value of $D_{P_F \to \mathcal{S}}$ is more than twice as high as for the state-dependent policies. This indicates that due to the increased complexity of the exchange mechanism a larger subpopulation size is necessary than for the state-dependent policy optimization. Also, the number of non-dominated solutions obtained for smaller subpopulation sizes for the exchange mechanism is further reduced than for the state-dependent policies. Compared to the solutions obtained for $n^{subpop} = 130$, the number of non-dominated solutions is reduced to about 63.2% and 77.5% for $n^{subpop} = 30$ and 80, respectively.

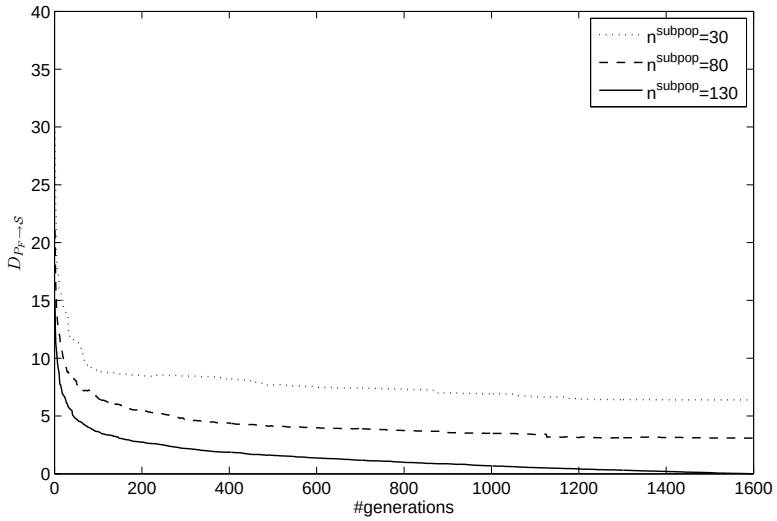


Figure 5.1: Convergence graphs for state-dependent allocation policies for varying subpopulation sizes

Similarly to the discussion in Section 4.5.4, the question arises what impact the final $D_{P_F \to \mathcal{S}}$ values for the different subpopulation sizes have on the quality of the obtained solutions. To address this issue we analyzed the

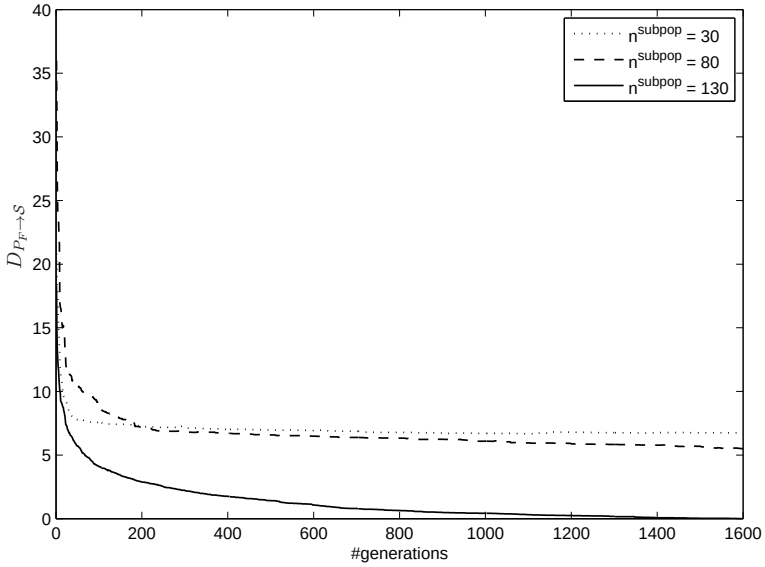


Figure 5.2: Convergence graphs for exchange mechanism for varying subpopulation sizes

obtained Pareto-fronts for the different population sizes. In Figure 5.3, and Figure 5.4 the results for the state-dependent policies and the exchange mechanism are presented for the different subpopulation sizes.

For both the state-dependent allocation policies and the exchange mechanism the difference in location of the obtained Pareto fronts between the different population sizes regarding is small. Solely, the range of the Pareto-fronts differs and the front obtained for $n^{subpop} = 130$ tends to extend to solutions with larger G_1 and G_0 values which are not reached using $n^{subpop} = 30$ or 80. For the state-dependent policies, the solutions using $n^{subpop} = 130$ extend further towards solutions with G_1 values larger than 140 for $G_2 > 100$. For the exchange mechanism, only the Pareto-front obtained using $n^{subpop} = 130$ include points with G_1 values greater than 150 which are not reached for the smaller population sizes.

Therefore, the choice of the subpopulation size mainly affects the resulting range of the Pareto-front in the G_1 dimension and the number of non-dominated solutions contained in the elitist archive of SDR-AVS-MIDEA. While the constricted range for the state-dependent policies mainly affects areas with large G_1 and G_2 values, the solutions obtained for the exchange mechanism for smaller n^{subpop} are limited to points with G_1 values smaller than 150. As the exchange mechanism outperforms the state-dependent policies and allocations optimized in Chapter 4 in this area, as we will see

in Section 5.5.4, $n^{subpop} = 130$ is imperative for the policy optimization of the exchange mechanism while a smaller subpopulation size can be sufficient for state-dependent policy optimization depending on the desired range and number of available non-dominated solutions.

Determining the required number of evaluations

Since the distance between the Pareto fronts obtained for the policy optimization appear small and the convergence graphs show a slower convergence after an increasing number of generations, we also evaluate whether decreasing the number of evaluations may result in comparable fronts to be obtained at reduced computational costs. Similarly to Section 4.5.3, we evaluate two approaches:

1. determine the number of generations needed to obtain comparable convergence performance, i.e. the $D_{P_F \to \mathcal{S}}$ value for $n^{subpop} = 130$ and 80 equals the $D_{P_F \to \mathcal{S}}$ value obtained for $n^{subpop} = 30$ value after 1600 generations,

2. determine the number of generations which have the same number of evaluations to be performed for $n^{subpop} = 130$ and 80 as for $n^{subpop} = 30$ and 1600 generations.

Using the first approach, the final $D_{P_F \to \mathcal{S}}$ value of 6.4 for $n^{subpop} = 30$ is achieved for $n^{subpop} = 80$ and 130 after 105 and 28 generations for the state-dependent policies, respectively. Considering the exchange mechanism, about 30 generations are required for $n^{subpop} = 130$. The results obtained following this approach show that the resulting Pareto fronts for the adaptive policies are very close to each other. However, we observed that the fronts contain considerably less non-dominated solutions than the approximated Pareto fronts obtained by the joint fronts over multiple runs of SDR-AVS-𝕄IDℍEA. As an example, the Pareto-fronts for the state-dependent policies for the different number of generations and subpopulation sizes are shown in Figure 5.5. Based on the data of Pareto optimal solutions we observed that the number of non-dominated solutions for $n^{subpop} = 80$ and 105 generations is decreased by almost 50% compared to the solutions obtained for $n^{subpop} = 30$ and 1600 generations. For $n^{subpop} = 130$ and 28 generations, the number of non-dominated solutions amounts to only a third compared with $n^{subpop} = 30$ after 1600 generations. In comparison to the results for 1600 generations, the fronts for $n^{subpop} = 80$ and $n^{subpop} = 130$ contain about 30% and 20% of the non-dominated solutions. Thus, we can conclude

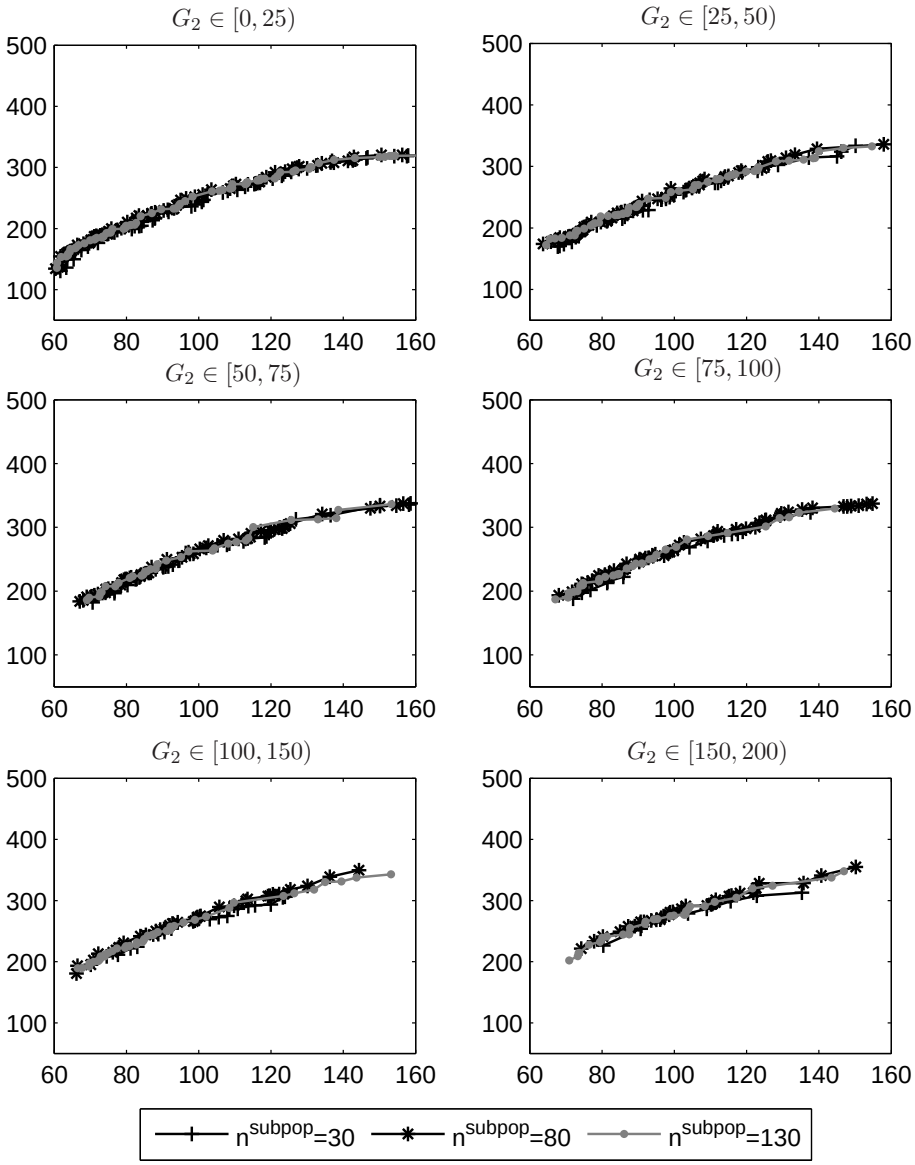


Figure 5.3: Pareto fronts for state-dependent allocation policies for varying subpopulation sizes after 1600 generations; x- and y-axes depict the corresponding G_1 and G_0 values, respectively

that the first approach results in a number of generations that is too small for obtaining comparably large and broad Pareto-fronts.

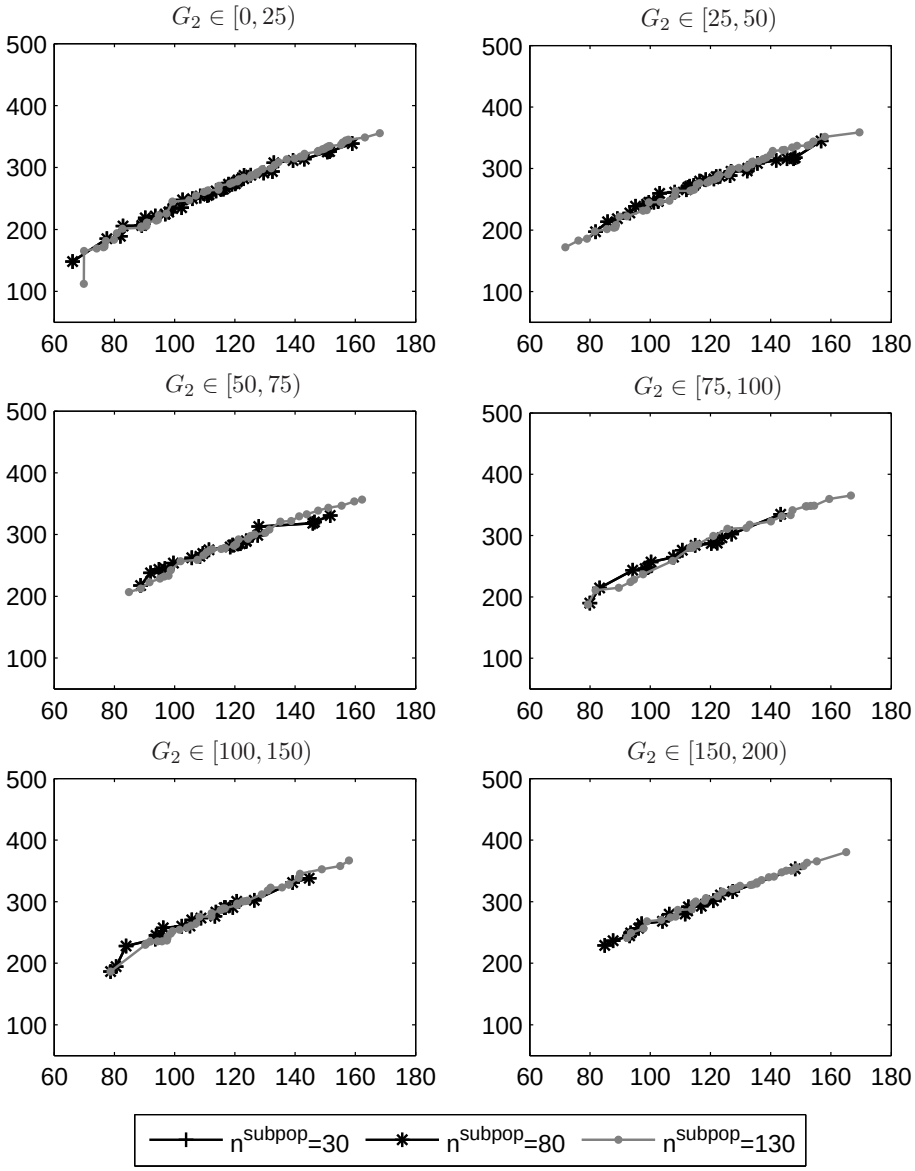


Figure 5.4: Pareto fronts for exchange mechanism for varying subpopulation sizes after 1600 generations; x- and y-axes depict the corresponding G_1 and G_0 values, respectively

Using the second approach, the number of evaluations involved in running SDR-AVS-𝕄IDEA using $n^{subpop} = 30$ for 1600 generations amounts to 134,520 which corresponds to about 600 and 370 generations for $n^{subpop} = 80$

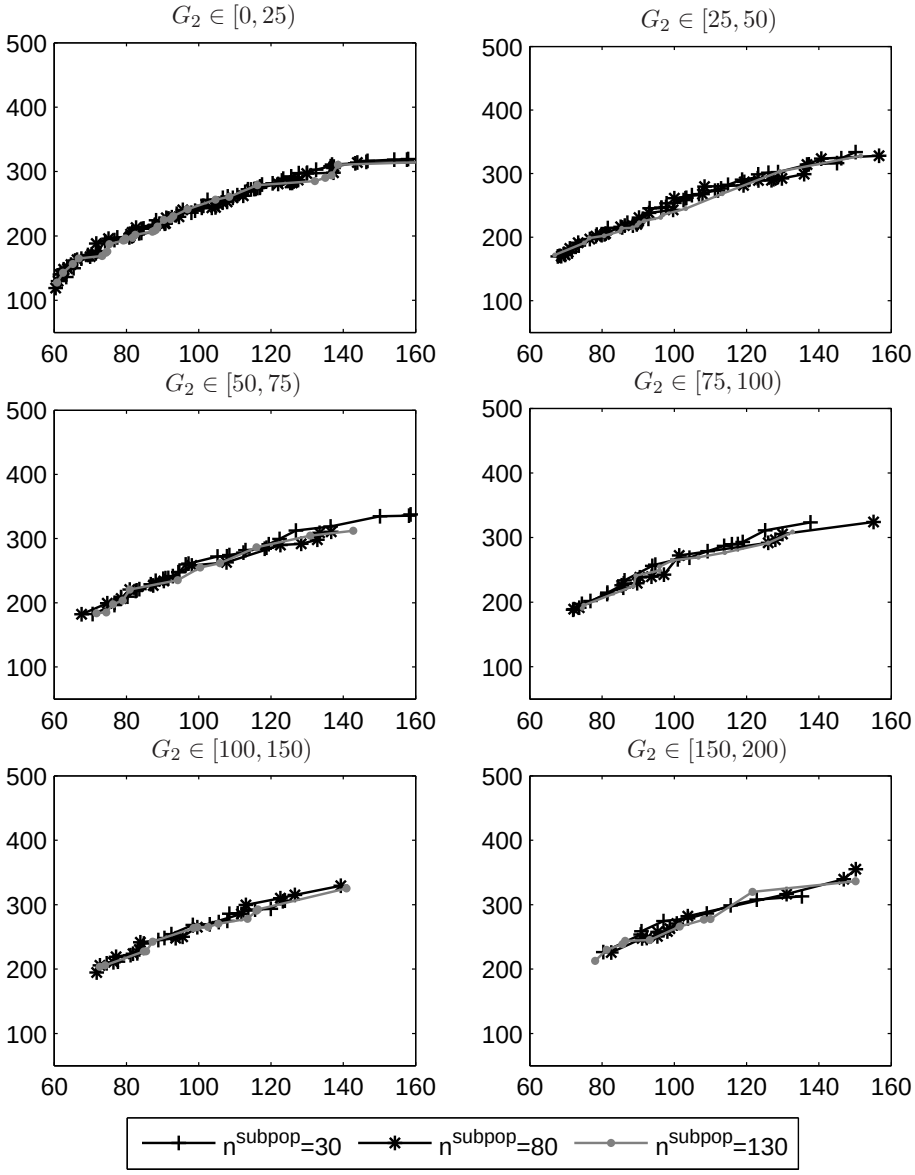


Figure 5.5: Pareto fronts for state-dependent policies after 1600, 105 and 28 generations for subpopulation sizes of 30, 80 and 130, respectively; x- and y-axes depict the corresponding G_1 and G_0 values, respectively

and 130, respectively. Comparably to the results in Section 4.5.3, we observed that this second approach results in a equivalent location of the ob-

tained Pareto fronts and an increase in the number of non-dominated solutions for $n^{subpop} = 80$ and 130, respectively. In comparison to the sets of non-dominated solutions obtained for $n^{subpop} = 80$ and 130 after 1600 generations, a decrease of about 35% can be observed, respectively. Therefore, we can conclude that the number of generations needed to obtain comparable optimization results can be reduced by a factor of 4 in the policy optimization.

The reduced number of evaluations results in a runtime of about 13 days when the multi-objective policy optimization is to be performed on a single PC. Since this optimization approach takes more than four times longer than the optimization presented in Chapter 4, the question arises whether optimized adaptive allocation policies result in an improved performance. To answer this question we analyzed and compared the obtained Pareto optimal solutions which is presented in the following section.

Intermediate conclusions Due to the fast convergence of SDR-AVS-MIDEA for the policy optimization problem, a considerable reduction in runtime can be achieved by reducing the number of generations with small losses in the number of non-dominated solutions compared to the initial setup of the optimization. For state-dependent policies, a further considerable reduction in runtime can be obtained through a smaller subpopulation size if the region of interest to the decision maker is restricted to smaller G_1 and G_2 values which also involves a further decrease in the size of the Pareto-optimal sets. For the exchange mechanism, however, a decreased subpopulation size limits the obtained Pareto front to G_1 values below 150 which is unfavorable to the applicability of the mechanism as will be discussed below.

5.5.4 Optimization results non-anticipatory policies

In order to provide a large range of optimized allocations, the following optimization results are obtained by 10 optimization runs of SDR-AVS-MIDEA using $n^{subpop} = 130$ and 370 generations. To crossvalidate our results, we also evaluated the obtained optimized allocation policies with 50 different random seeds. In this section we first evaluate the optimization results obtained for the non-anticipatory allocation policies and analyze the optimized parameter settings. For benchmarking the performance of the non-anticipatory policy optimization approach, we also include the results optimized in Chapter 4 as a benchmark.

Analysis of the obtained Pareto fronts

Figure 5.6 shows the Pareto fronts obtained for the different allocation policies and the static allocations optimized in Chapter 4. The results show that the state-dependent policies outperform the exchange mechanism for G_1 values below about 140, depending on value of G_2. For larger G_1 values, however, the exchange mechanism provides for greater patient throughput within the same bounds for the back-up capacity usage.

Considering also the static allocations optimized in Chapter 4, the state-dependent policies outperform the static solutions, followed by the exchange mechanism for values of G_1 below 100 to 125, depending on the G_2 setting. For G_1 values between about 100 and 140, depending on the values of G_2, the bed exchange mechanism performs second-best after the state-dependent adjustments and surpasses the state-dependent policies for G_1 values of approximately 130 and higher. The reason why the exchange mechanism shows a high performance for larger G_1 values is that the mechanism requires larger base resource allocations in order to be able to shift resources between the units. For lower resource costs the interaction between the hospital units provides that required allocation adjustments cannot be sufficiently undertaken using solely resources that are already allocated to the system of care units which explains the decline in performance.

Although the visual difference between the different policies appears rather small in Figure 5.6, the reader should note that the results are based on simulation runs of 4 weeks which means that a small difference achieved in this period will become more apparent when considering the performance over a longer time period.

The crossvalidation evaluations are comparable to the EDA results for the state-dependent allocation policies as depicted in Figure 5.7. For the exchange mechanism, the crossvalidation results even show a somewhat higher performance than the results obtained by the EDA (Figure 5.8). Thus, the exchange mechanism is likely to perform better yet in a practical setting.

Analysis of optimized policy parameters

For analyzing the optimized parameters of the different adaptive policies and static allocation parameters determined in Chapter 4, we use boxplots to visually summarize the results and statistically assess the differences between the solutions using the analysis of variance (ANOVA) technique for multiple response variables (MANOVA) and multiple ANOVA for individual dependent variables [50].

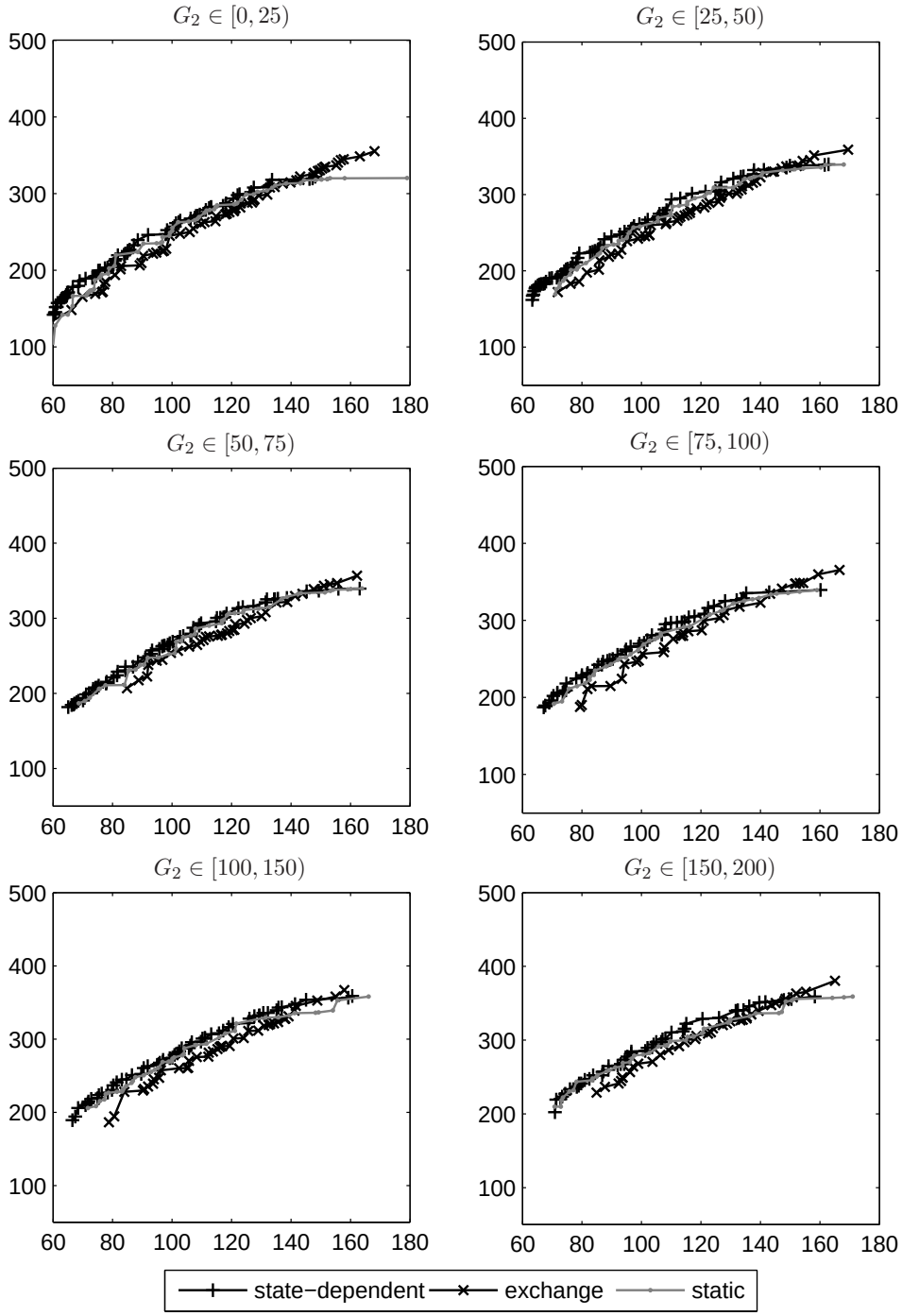


Figure 5.6: Pareto fronts for adaptive policies including static allocations optimized in Chapter 4; x- and y-axes depict the corresponding G_1 and G_0 values, respectively

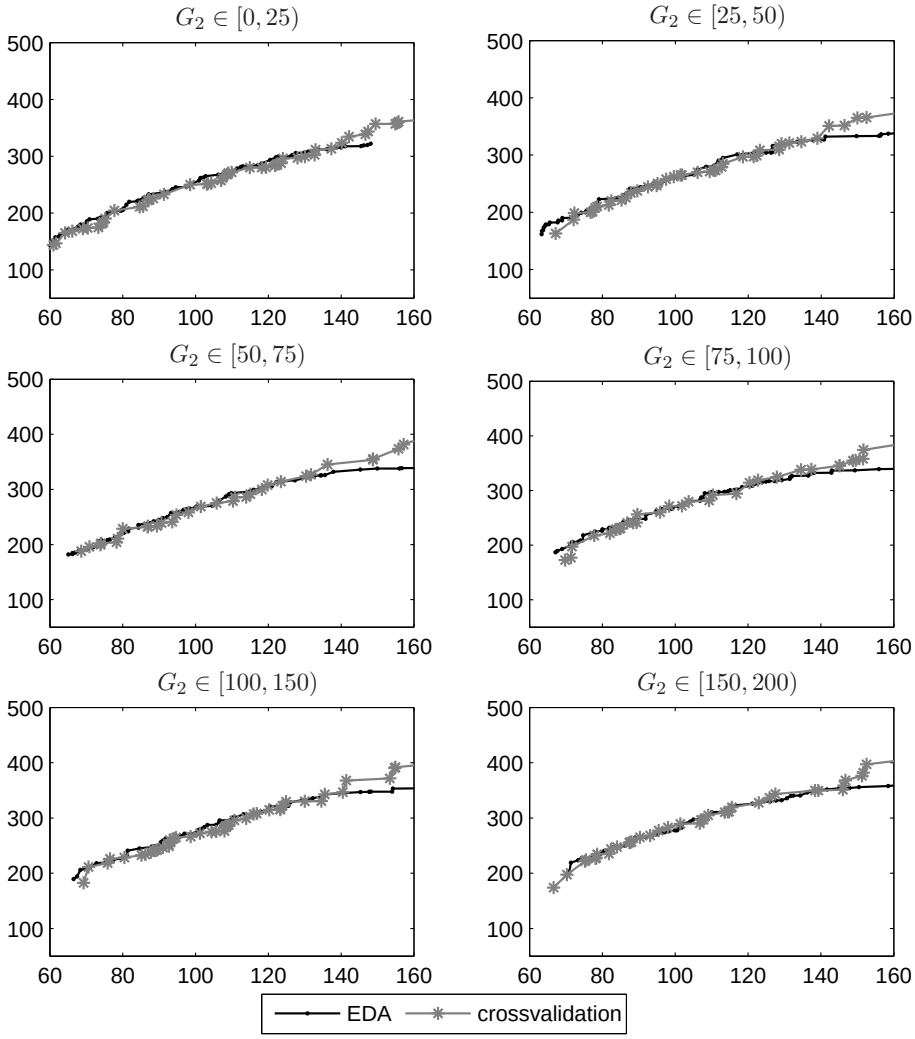


Figure 5.7: Pareto fronts for state-dependent allocation policies including crossvalidation results; x- and y-axes depict the corresponding G_1 and G_0 values, respectively

Using MANOVA, we can conclude that the different base allocation parameters as a group are significantly different for the optimized static allocations in Chapter 4 and the adaptive policies proposed in this chapter (at a level of significance $\alpha = 0.05$). The respective parameters are shown using grouped boxplots in Figure 5.9. Based on visual inspection we note that the base allocations for the exchange mechanism are higher than the state-dependent base and static allocations for the IC, IC-HC and MC and

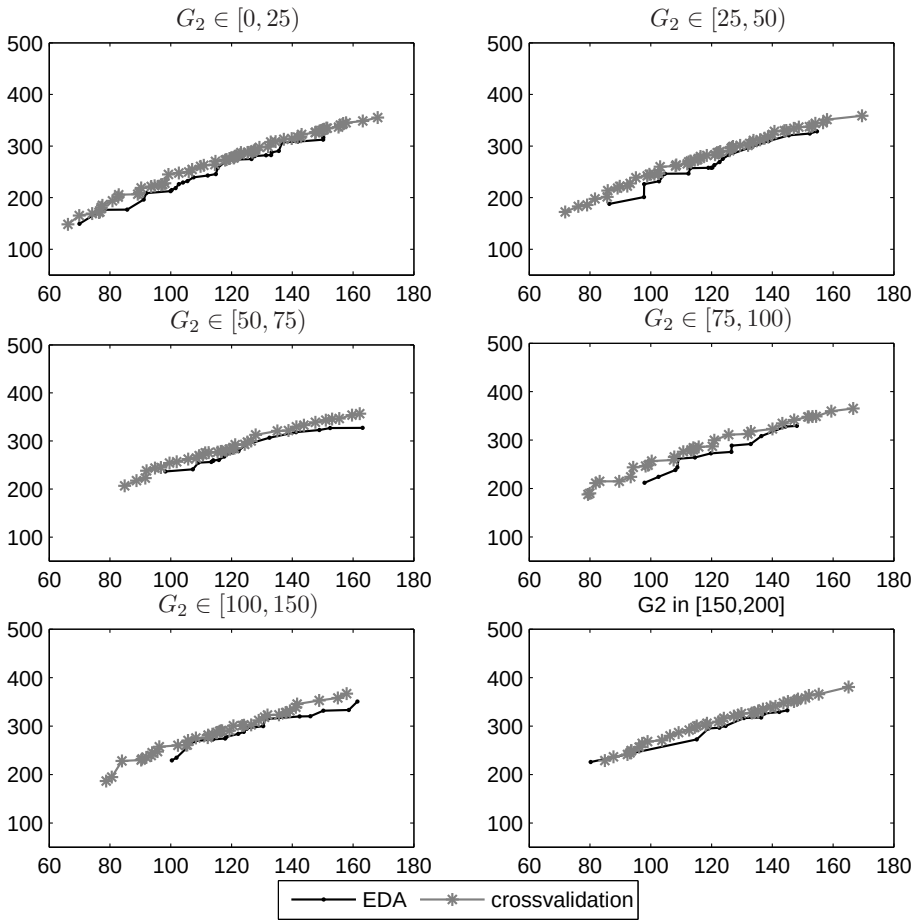


Figure 5.8: Pareto fronts for exchange mechanism including crossvalidation results; x- and y-axes depict the corresponding G_1 and G_0 values, respectively

comparable for the CTS-OR and CTS-HC. For the CTS-PACU and CTS-ward, the base allocation for the exchange mechanism appears to be smaller than for the optimal fixed allocation and state-dependent base allocations, respectively. Moreover, the range of base allocations for the exchange mechanism appears smaller. Using the ANOVA technique, the higher base allocation of the exchange mechanism appears to be significant, also for the CTS-OR. For the CTS-PACU the difference is not significant ($p = 0.06$). The larger base allocation is a logical consequence of the resource exchange mechanism. Since a larger allocation is a precondition for enabling resource exchange between different units and care levels according to Algorithm **??**,

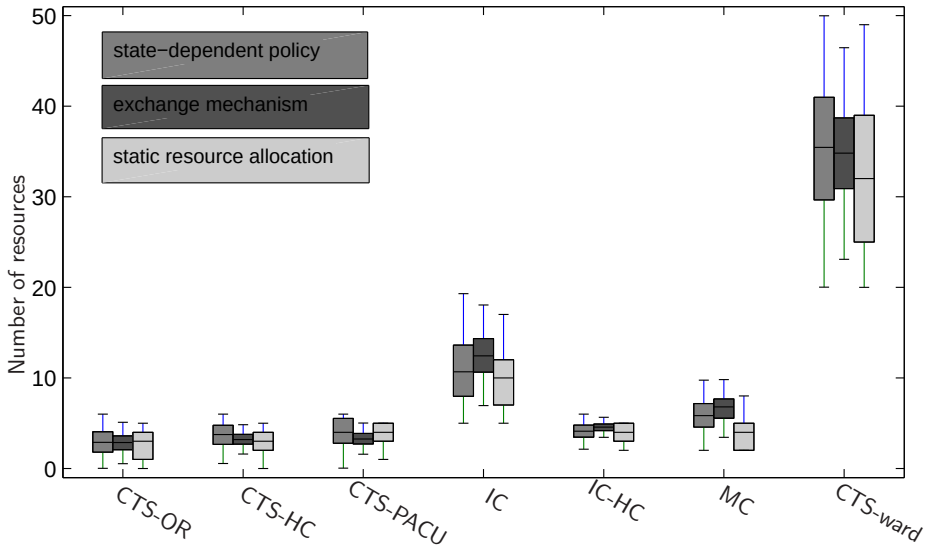


Figure 5.9: Boxplots of optimized base allocation parameters, r_u^{base}, including static allocations optimized in Chapter 4

this steers the search using SDR-AVS-𝕄IDEA towards larger allocations. Specifically, a significantly larger base allocation of level 1 and 2 resources is also essential for facilitating resource adjustments to all three care levels since upgrading of resources is not permitted, cf. Section 5.4.2. The smaller base allocation for the state-dependent policies can also be explained by the unrestricted possibility to add and remove resources when indicated by the allocation policy.

Allocation parameters	**CTS-OR**	**CTS-HC**	**CTS-PACU**	**IC**	**IC-HC**	**MC**	**CTS-ward**
$r_u^{base} = r_u^{min}$	4.2%	1.17%	2.16%	4.48%	5.26%	2.89%	1.15%
$r_u^{base} = r_u^{max}$	5.63%	8.37%	26.97%	0.65%	6.72%	1.85%	1.09%
Online allocation							
$r_u(t) = r_u^{max}$	n.a.	54.81%	32.12%	35.07%	18.99%	7.19%	13.92%

Table 5.2: Proportion (%) of optimal solutions that take a base allocation parameter value r_u^{base} equal to the allocation bounds r_u^{min} and r_u^{max} and frequency of maximal allocation online during simulation through resource adjustments for state-dependent resource allocation policies

Table 5.2 and Table 5.3 summarize the proportion of obtained solutions that feature extreme base allocation parameter values or maximal allocations during simulation for the state-dependent policies and the exchange mechanism, respectively. We can thus conclude that only a small proportion of the optimized policies use a minimal base allocation. Moreover, we observed that a minimal allocation at CTS-OR, CTS-HC and CTS-PACU occurs for only 0.8% and 0.1% of the state-dependent policies and the exchange mechanism, respectively.

The maximal allocation is used but rarely as base allocation in the optimized state-dependent policies, except for the CTS-PACU where a maximal allocation is used in more than 1 out of four policies. The exchange mechanism features fewer optimized solutions with maximal base allocations for CTS-OR and CTS-PACU and a comparable frequency for the CTS-HC, but more solutions with maximal base allocations at IC, IC-HC, MC and CTS-ward which can be attributed to the adjustment restrictions and the possibilities for resource exchange, cf. Section 5.4.2 and Section 5.5.2. During simulation, the state-dependent policies frequently use the maximally allowed resource allocation, especially at the CTS-HC and IC. Applying the exchange mechanism, the maximal allocation is used significantly less often. Since the use of the maximally allowed resources involves large investments for hospital management, the exchange mechanism can thus be implemented more easily in practice.

Allocation parameters	**CTS-OR**	**CTS-HC**	**CTS-PACU**	**IC**	**IC-HC**	**MC**	**CTS-ward**
$r_u^{base} = r_u^{min}$	6.61%	3.36%	0.67%	3.36%	1.12%	5.38%	0.89%
$r_u^{base} = r_u^{max}$	4.26%	8.07%	3.36%	2.8%	16.93%	8.52%	6.73%
Online allocation							
$r_u(t) = r_u^{max}$	n.a.	1.58%	n.a.	0%	0.57%	0.52%	0.01%

Table 5.3: Proportion (%) of optimal solutions that take a base allocation parameter value r_u^{base} equal to the allocation bounds r_u^{min} and r_u^{max} and frequency of maximal allocation online during simulation through resource adjustments for exchange mechanism

Moreover, we analyzed the frequency of performed resource adjustments using the different adaptive allocation policies. From Table 5.4 we can conclude that the resource allocations under state-dependent allocation policies are relatively variable and that the added interaction and constraints for shifting resource capacity from one care level to another, considerably decreases the frequency of resource adjustments under the exchange mecha-

nism. Thus, the exchange mechanism can be implemented in practice more easily as fewer resource adjustments also require less flexibility of the staff.

Frequency	**CTS-HC**	**IC**	**IC-HC**	**MC**	**CTS-ward**
state-dep. policy	14.8%	6.63%	14.08%	17.99%	15.07%
exchange mech.	4.96%	0.06%	4.75%	5.71%	0.93%

Table 5.4: Frequency (%) of resource adjustments performed for adaptive resource allocation policies

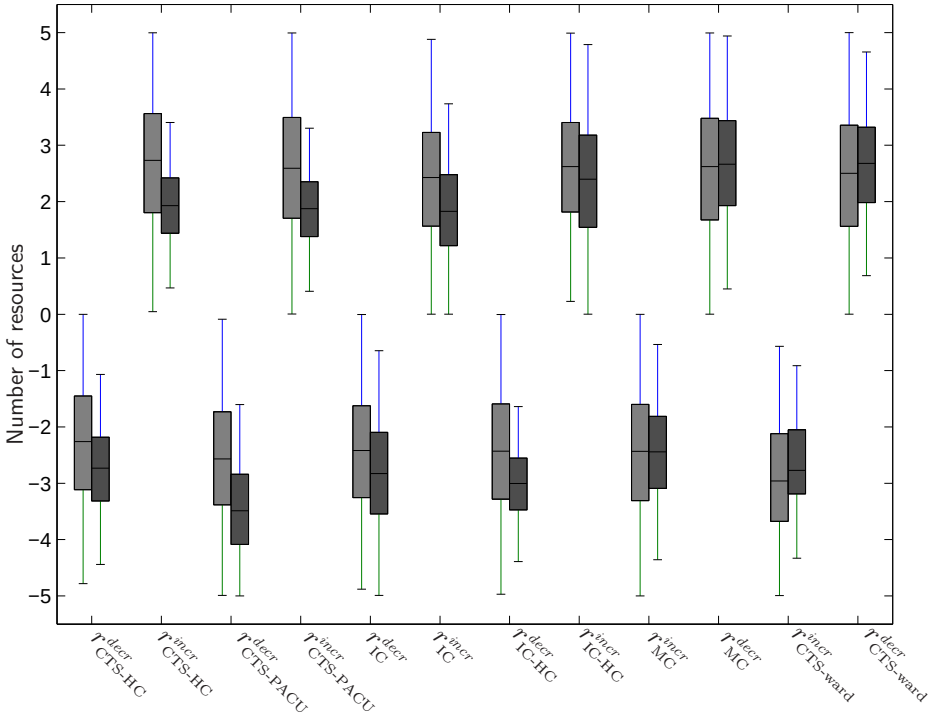


Figure 5.10: Boxplots of optimized resource adjustment parameters, r_u^{decr} and r_u^{incr} where parameters of state-dependent policies and exchange mechanism are depicted by grey and dark-grey bars, respectively

In Figure 5.10 the in-/decremental adjustment parameters are summarized using grouped boxplots for the optimized state-dependent policies and exchange mechanism policies. We can observe that the incremental adjustment parameters for the state-dependent policies appear higher than for

the exchange mechanism, while the decremental adjustments are smaller. This finding is also confirmed to be significant using ANOVA, except for the incremental adjustment at MC ($p = 0.07$, $\alpha = 0.05$). This result can be expected because the exchange of resources has an impact on the entire network of care units and a great variation in resource allocation can be disturbing. Since an in-/decrease in resource capacity depends on the resource availability at the other units, the policies provide for smaller adjustments in order to prevent an imbalance of resource occupancy in future time periods. This is implicitly learned and measured by the EA during the simulation period. Together with the exchange being the result of the resource availability at the different units in case of the exchange mechanism, this means that the allocated resource capacity is significantly more volatile if state-dependent policies are employed.

The thresholds used in the adaptive policies for the exchange mechanism and their range are significantly higher for the CTS-ward and the IC-HC than for the state-dependent policies, whereas the opposite holds for the thresholds and their range for the IC and MC. This means that the policies for resource availability at CTS-ward and IC-HC tend to reduce resource capacity rather than to increment it. For the IC and MC, more frequent adjustments are provided through the smaller range of utilization thresholds.

Figure 5.11 depicts the patient throughput resulting from the optimized allocation policies for the different patient groups. We can conclude that both the adaptive allocation policies feature a heterogeneous patient mix. The state-dependent policies consistently achieve the highest patient throughput for all patient types. Compared to the mix achieved through the static resource allocations optimized in Chapter 4, the exchange mechanism solely achieves a greater throughput for type III and IV patients. For type I+II patients, the average throughput over all the obtained Pareto optimal exchange mechanism solutions is smaller than for the static allocations, but also shows a higher lower bound. The actual throughput depends on the chosen Pareto optimal solution.

We can observe that due to the larger variability in the number of allocated resources, the variability in patient throughput is the highest for the state-dependent allocation policies. Since the exchange mechanism parameters feature a smaller range of base allocations and smaller and fewer adjustments, the range of patient throughput achieved by the exchange mechanism is comparatively small.

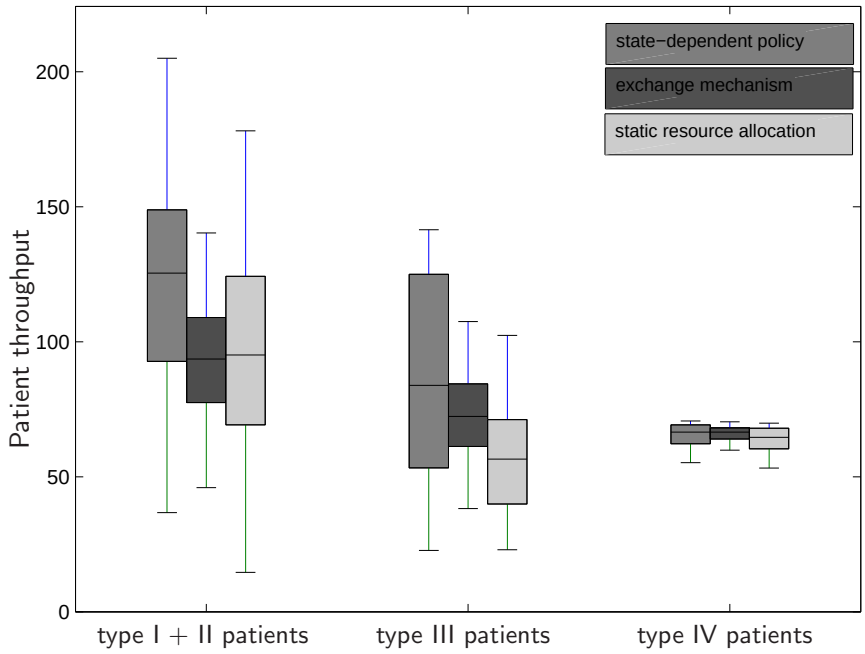


Figure 5.11: Boxplots of patient throughput resulting from optimized adaptive allocation policies including static allocations optimized in Chapter 4

Intermediate conclusions We can conclude that the state-dependent allocation policies and the exchange mechanism differ considerably in their policy parameters and the resulting online allocation. While the state-dependent policies provide for frequent maximal allocations, the additional constraints imposed on the resource allocations in the exchange mechanism provide for fewer and smaller resource adjustments. Due to the large investments required for increasing the available resources, the frequency of maximal allocations is an important factor to be taken into account when choosing an adaptive allocation policy. Moreover, the proposed MO policy optimization approach provides for a heterogeneous patient mix that is considerably higher for the state-dependent policies due to the high flexibility in allocating resources. However, the increased variability in resource allocations also provides for more variability in the resulting patient mix compared to the exchange mechanism and the static allocations optimized in Chapter 4. The actual throughput depends on the chosen Pareto optimal solution.

5.5.5 Optimization results anticipatory allocation policies

Furthermore, we evaluated the impact of anticipation in the optimization of resource allocation policies. Predicting the future resource occupancy using forward simulation involves increased computational costs. Running a single generation takes about one hour where the policy evaluations are performed in parallel on 24 nodes of the high-performance computer cluster used for the experiments in this chapter. This is also the reason why the results presented in this section are restricted to 4 runs of SDR-AVS-𝕄IDEA for 100 generations. The small number of evaluations in combination with the increased complexity of the optimization provide for the smaller number of non-dominated solutions obtained for this approach. As the convergence behavior for state-dependent policies depicted in Figure 5.1 show that 90% of the overall convergence are achieved within the first 100 generations, the results can be considered as sufficiently indicative.

Does the inclusion of prediction information improve the performance of state-dependent policy optimization?

For the allocation problem in the above setting, the direct use of predicted future occupancy information in the state calculation appeared not to improve the results obtained by using current occupancy information and daily resource adjustments. The corresponding two-dimensional Pareto-fronts are depicted in Figure 5.12. This result indicates that the optimization using state-dependent allocation policies as defined in Section 5.4.1 in combination with SDR-AVS-𝕄IDEA is able to exploit the best allocation possibilities for the treatment and arrival processes in the model. A possible explanation could be that the design of the allocation policies is well aligned with the problem at hand. The design then potentially provides for the inherent anticipation of time-dependence effects in the optimization of the policy parameters. Due to the fact that the simulation is run for a certain period of time, time-dependence effects are incorporated in the fitness evaluation. Thus, the optimization of adaptive policies using SDR-AVS-𝕄IDEA and the simulation already appears to be an effective approach that inherently takes future consequences of an allocation decision into account.

When could prediction information improve the performance of state-dependent policy optimization?

The added-value of anticipation becomes substantial if the model of the patient arrival processes is extended such that an increased availability of

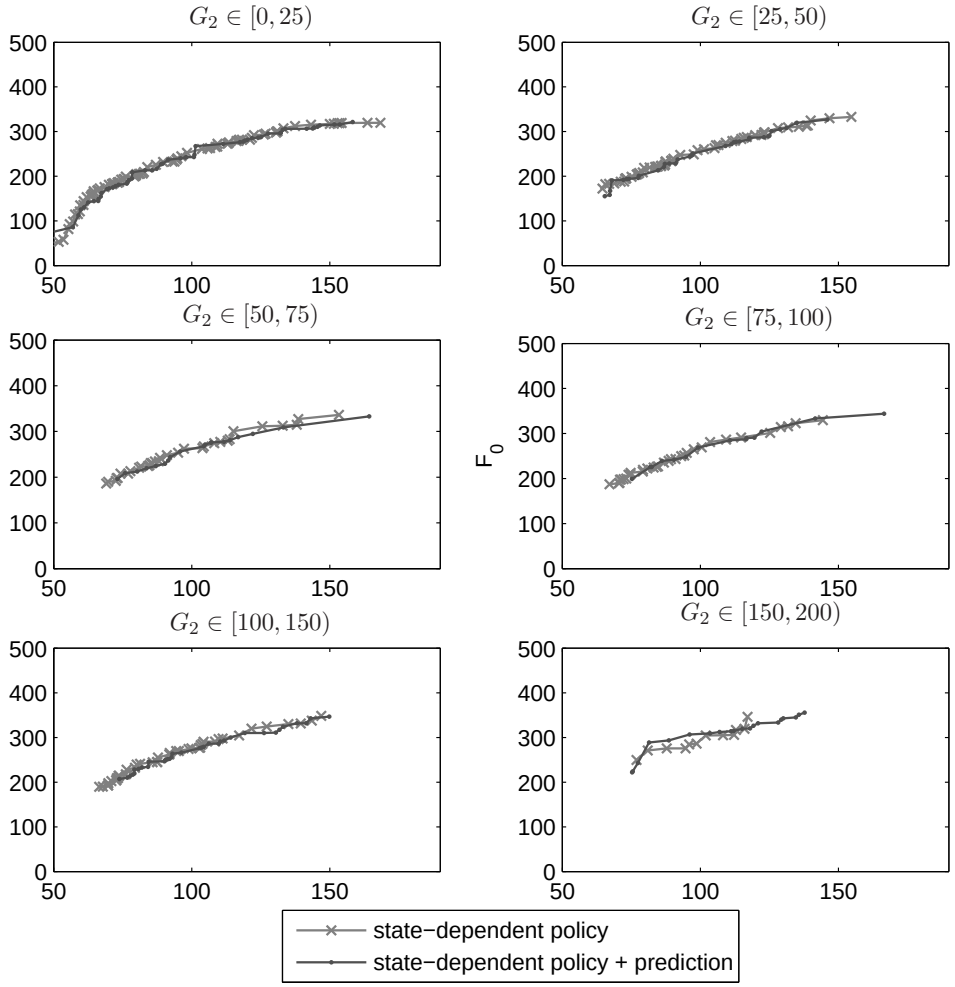


Figure 5.12: Pareto fronts for state-dependent policies with an adjustment period of 2 days using current and predicted future occupancy information without adaptive patient arrivals (x- and y-axes depict the G_1 and G_0 values, respectively)

resource capacity entails an increase in demand of this type of care or treatment. This extension models the notion of an enlarged attraction of patients that follows a physical capacity enlargement or an increase in competence or reputation of a hospital division or specialty. We assume an adaptive demand model such that an additional bed at the admitting unit increases the corresponding current mean daily demand by factor 2, while a removed bed decreases the demand by the inverse. Demand for type I + II patient admissions is not affected since we assume sufficiently long waiting lists for

type I and II patients, cf. Chapter 2. Thus, this demand model affects the number of probes for type III and IV patient admissions requested by others. Consider for example probes for type III patient admissions by ambulance services. Ambulance services may be more likely to request the admission of an emergency patient if the IC capacity of a hospital is increased since the capacity increase also enhances the chances that the requested admission is accepted. An increase in type IV patient admission requests might be resulting from more frequent referrals from general practitioners or recommendations from health insurance companies due to a decreased waiting time for elective surgery if more capacity is available.

In order to limit a possible imbalance in the patient mix, the demand was limited by 20 times the initial mean demand, cf. Section 2.4.1. The admission policy remains unchanged as described in Table 2.2 on page 46.

Analysis of the obtained Pareto fronts In Figure 5.13 the Pareto fronts computed over all runs are shown for optimizing the adaptive policy without and with predicted state information. Because the interpretation of three-dimensional Pareto fronts can be hard, the results are also presented by computing the Pareto front for only the first two objectives, similarly to the representation above. The third objective, G_2, is categorized into three ranges $[0, 25]$, $[25, 50]$ and $[50, 75]$ resulting in the presentation of three Pareto fronts in Figure 5.14 for increased levels of back up capacity usage.

The difference between the results is substantial. Without prediction included in the policy, better results than the current real-world results can still be obtained. With prediction however, the MOEDA picks up on the fact that increased availability of resource capacity entails an increase in demand for care of this type of treatment. The performances of the state-dependent policies with and without prediction are comparable for small resource costs ($G_1 \leq 70$). For larger resource costs the use of predicted occupancy information in the state-dependent policy considerably improves the throughput. An increase in the total throughput of 200% and more can be observed. The increased frequency in demand for care that is established by the optimized policies using prediction also results in more efficient usage of beds in the sense that there are now few to no results for higher rates of backup-capacity usage. Policies that lead to high use of backup capacity are dominated by polices that have low backup-capacity usage because a high throughput is possible even with low backup capacity usage. This can partly be attributed to the fact that an available bed can be occupied by a new patient much faster due to the increased admission requests. To obtain

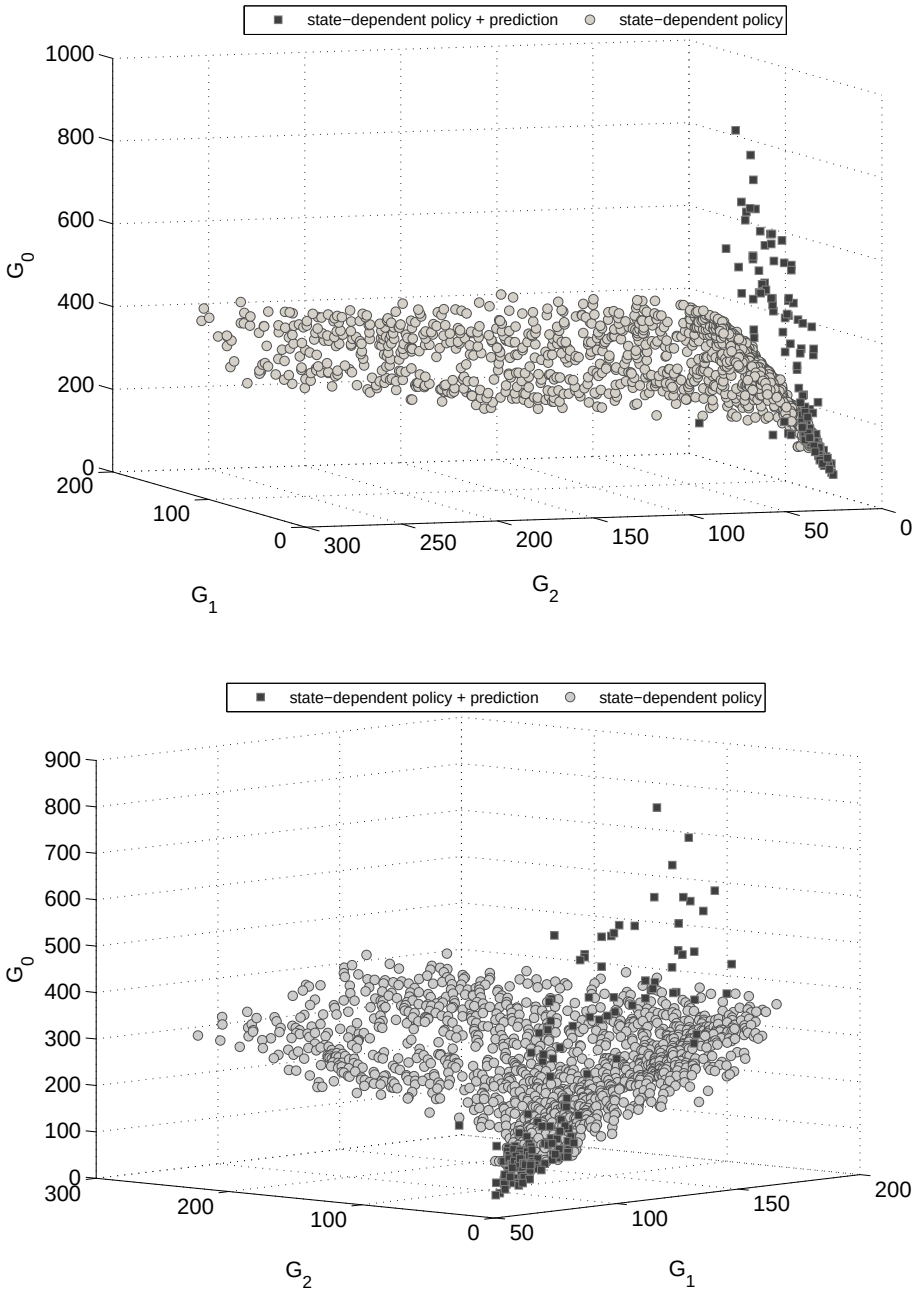


Figure 5.13: Three-dimensional Pareto front obtained from multiple runs from two perspectives

such results, however, prediction of future resource usage is required. The crossvalidation results are comparable with the EDA results with a slight decrease in patient throughput for high resource costs and little back-up usage ($G_1 \geq 120$ and $G_2 \leq 25$).

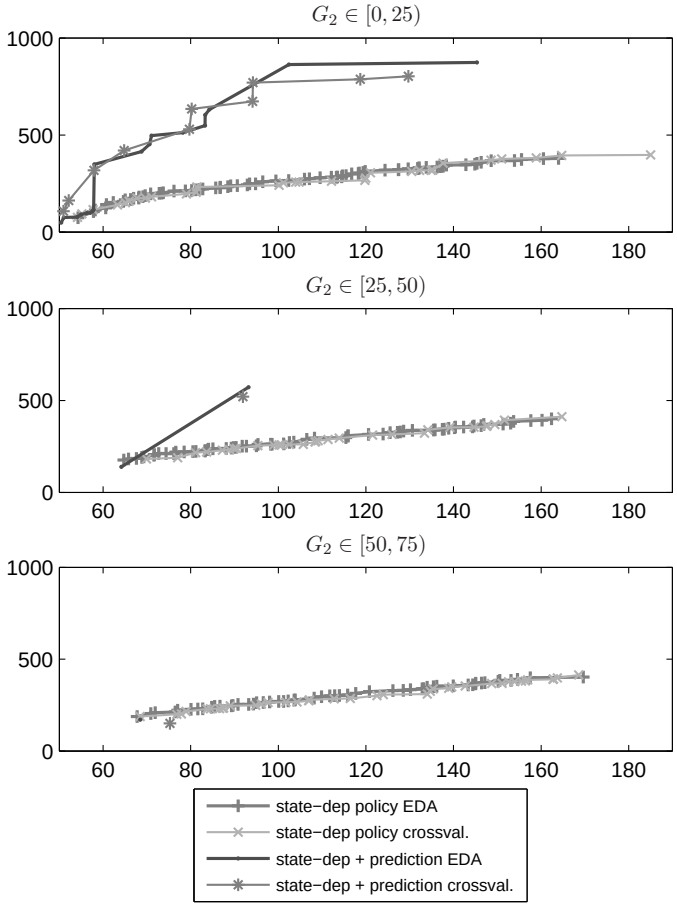


Figure 5.14: Pareto fronts for state-dependent policies with an adjustment period of 2 days using current and predicted future occupancy information including crossvalidation results; x- and y-axes depict the G_1 and G_0 values, respectively

Analysis of optimized policy parameters and resulting patient flows In Figure 5.15 grouped boxplots of the base allocation parameters are depicted for the (non-)anticipatory policies and the extended patient demand model. We can observe that the parameters for the anticipatory policies are considerably higher for the CTS-ward, the IC and the CTS-OR.

For the CTS-HC and CTS-PACU, the bullk of the base allocations for the anticipatory policies is lower than for the non-anticipatory policies and for the remaining units the base allocations appear comparable. For the CTS-ward, IC, CTS-PACU and CTS-OR the differences are significant which is determined using multiple ANOVA tests at $\alpha = 0.05$.

Using ANOVA, we observed that the in- and decremental adjustment parameters do not significantly differ between the (non-)anticipatory policies. Rather, a significant discrepancy can be noted in the utilization thresholds that determine whether the current allocation should be adjusted. Specifically, the lower and upper thresholds, UT_u^{decr} and UT_u^{incr}, for MC, CTS-PACU and CTS-HC are significantly smaller for the anticipatory policies, while the adjustment thresholds are significantly higher for the IC-HC and IC which are the care units for which patient demand is coupled to the available capacity. Thus, anticipation here improves the decision making through setting better thresholds and while providing for comparable resource adjustments.

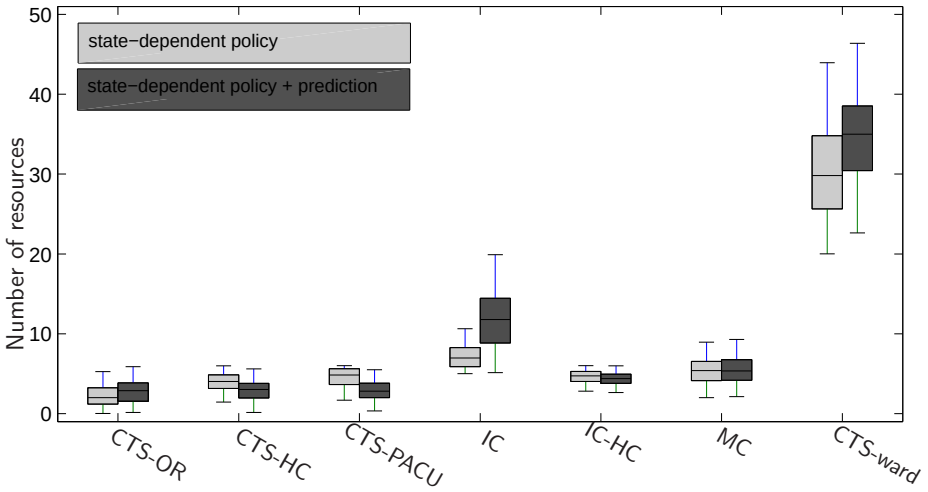


Figure 5.15: Boxplots of optimized base allocation parameters, r_u^{base}, for (non-) anticipatory policies

The patient mix resulting from applying the anticipatory policy is summarized using boxplots in Figure 5.16. We can observe that in general the resulting patient flows tend to a higher number of treated type III and IV patients and a slightly decreased type I+II throughput. The patient mix is still heterogeneous as an omission of type I+II patients occurs for only 0.8%

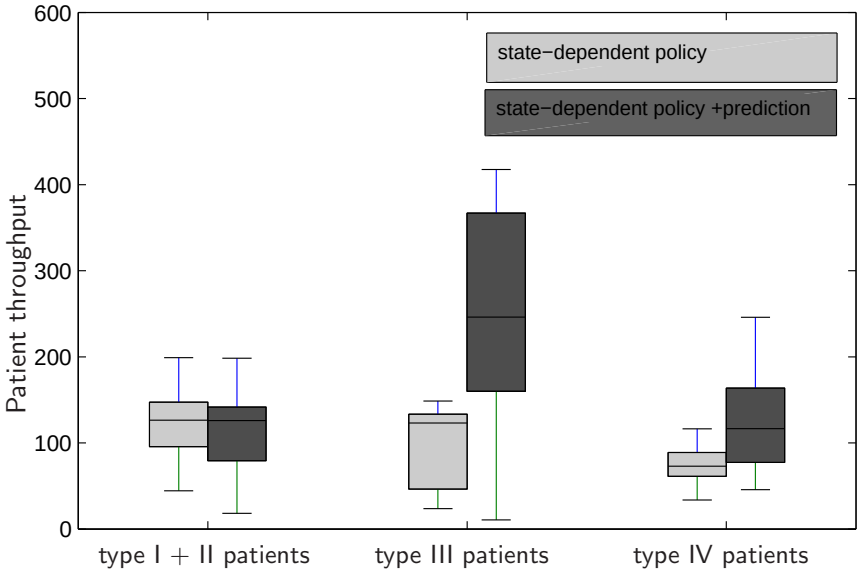


Figure 5.16: Boxplots of patient throughput resulting from optimized anticipatory allocation policies

of the policies. A throughput that is comparable to the CHE case is achieved by only about 2% of the policies which can be attributed to the extended demand model. Moreover, the throughput for all patient types shows a larger variation for the anticipatory policies compared to the state-dependent policies where no prediction information is used in the state calculation. Also, the skewed distribution of the type III and IV patient throughput is to be noted.

Intermediate conclusions We can conclude that the use of anticipatory policies does not affect the in-/decrement of the allocation. Rather anticipation provides for better aligned adjustment thresholds which result in a considerably improved performance. Moreover, the proposed MO optimization approach provides for a heterogeneous patient mix that features a higher proportion of type III and IV patients in the overall mix which can be attributed to the coupling model and admission policies for the different patient types in the system. It should be noted that the adaptive demand model and the obtained results may be somewhat exaggerated for the CHE situation. However, the extension reflects hospital reality where resource availability affects the frequency of admission requests and our results illus-

trate the possible contribution of anticipation in dynamic multi-objective decision making.

5.6 Conclusions

In this chapter we presented a multi-objective optimization approach for adaptive hospital resource management. We present a policy optimization approach for which we designed policies that allow for the dynamic allocation of resources in a network of care units. Due to the complexity of the allocation policies and the dynamic application domain, we used a state-of-the-art evolutionary MO technique, SDR-AVS-𝕄IDEA. The design of the policies that are evaluated online in the simulation allows an offline evaluation of the policies determined using the realistic and complex simulation described in Chapter 2. In our experiments we analyzed the convergence behavior of SDR-AVS-𝕄IDEA for different population sizes and determined the minimally required number of evaluations in the optimization which reduces the runtime of a policy optimization run by factor 4. Moreover, our results showed that adaptive policies can improve the optimized allocations presented in Chapter 4, and the design of policies allows the policies to be easily understandable for hospital experts which facilitates the implementation in hospital practice. The exchange mechanism presented in this chapter enables the actual implementation of the adaptive policies in practice without great changes in the current way of working or large investments in additional resource availability. Furthermore, we showed that policies that incorporate predicted information of future consequences of an allocation decision result in further improvements in the extended patient demand model. The improvements in performance are made possible by the design of the policies. SDR-AVS-𝕄IDEA then is powerful enough to detect and exploit the additional possibilities. In the original model, using SDR-AVS-𝕄IDEA in combination with the adaptive allocation policies and the simulation already appear to be an effective approach that inherently takes future consequences into account due to the time dependence effects incorporated in the simulation. Our results demonstrate that proper design in combination with state-of-the-art EAs given a sufficient population size can make an important contribution and achieve an improvement for complex real-world dynamic MO problems as in hospital resource management. An additional advantage of our policy types is that offline MO techniques can be used to optimize the parameters of the allocation policies. Furthermore, the resource allocation policies obtained by our approach feature a

heterogeneous patient mix. The actual patient mix depends on the Pareto optimal solution which may be an important issue for hospital management to consider when choosing an allocation policy to be implemented.

The proposed approach is very flexible as the model parameters can be easily adjusted to different hospital settings and a single allocation policy can be applied to the different scenarios in the simulation. Moveover, the policies were developed in cooperation with domain experts which provides for the policies being easily understood by hospital professionals which is of importance for the implementation and understanding in hospital practice.

It should be noted that when implementing the anticipatory policies in hospital practice, forward simulation is still required for the decision process because the prediction information is used as input for the allocation policy. Although the optimization of the policy parameters is the most time-consuming factor, which only has to be performed once, forward simulation needs to be performed at the beginning of each working day in order to decide upon the allocation decision to be taken. For this aim, the simulation would need to be synchronized with the current state at the real-life hospital units. Alternatively, learning techniques like neural networks, as considered in Section 3.5, may be employed to reduce the required computational resources.

Including prediction in the dynamic allocation policies for resource management is beneficial in the case where the model allows for adaptation of patient arrivals to the available resource capacity. The rationale behind this assumption is that increasing the available capacity provides for a higher chance of emergency admission and/or shorter waiting times for elective admissions that attract additional patients. A hospital might be contacted more frequently concerning emergency admissions by ambulance services due to an increased chance of admission. Also, shorter waiting times at hospitals have a pull effect on patient demand as patients increasingly include waiting times in their choice of health care service provider. Moreover, health insurance companies offer mediating services and try to find alternative hospitals with short waiting times to accelerate the patients' admission for clinical treatment. Furthermore, an increased resource capacity may contribute to raise the reputation of a hospital specialty.

It should be noted that the coupling of the arrival processes to the resource capacity is modeled in an abstract way. A meticulous model of the adaptive patient demand for a specific hospital setting would typically involve a large-scale economic analysis of the relevant patient groups in a multi-hospital setting which is beyond the scope of this thesis. For this reason, the effect in our model may be somewhat exaggerated, which was

done for the purpose of illustrating the contribution of anticipation in solving the dynamic optimization problem. For this reason, the final results for adaptive resource allocation policies using predicted information may be overly optimistic for the CHE problem where the coupling of the arrival processes and the resource allocation is probably weaker and less instantaneous. Further modeling of the underlying coupling in a multi-hospital setting is needed to assess the extent of improvement through anticipation. However, our results illustrate that our approach successfully allows to tackle the time-dependence problem for dynamic optimization problems under multiple objectives using our designed allocation policies.

Chapter 6

Discussion and conclusions

Planning decisions concerning patient flow logistics in hospitals are often taken in a decentralized way. This means that different specialized hospital units decide autonomously on e.g. patient admissions and schedules of shared resources. In this thesis we presented methods and techniques that provide decision support in this setting. In our approach we combined techniques from multi-agent systems and computational intelligence which allowed us to consider the dynamics of the problem while reflecting the distributed decision-making practice in hospitals. Specifically, we have designed and analyzed computational methods for (adaptive) hospital resource management, the prediction of future resource occupancy and the application thereof.

Our agent-based model captures multiple hospital care units and their decision policies, multiple patient groups with stochastic treatment processes and uncertain resource availability due to overlapping patient treatment processes. We have developed a simulation for the agent-based model and demonstrated its usefulness for decision support in this setting. Moreover, we applied learning and optimization techniques from computational intelligence (CI). We studied the use of CI prediction methods to predict future hospital resource occupancy resulting from admission and allocation decisions and analyzed the underlying probability distribution. Moreover, the applied CI techniques allowed us to design and evaluate improved (adaptive) decision policies for the agent-based model. We showed that the benchmark allocations obtained from the case study could be considerably improved using the multi-objective evolutionary optimization approach presented in this thesis. Furthermore, we showed that adaptive decision policies and the inclusion of predicted resource usage resulted in further improvements.

Moreover, the decision policies can be implemented in hospital practice in a straight-forward manner because the agent-based model closely resembles the real-world situation.

In the remainder of this chapter, we reflect on the applicability of our approach in a real-life hospital setting, as well as on future research possibilities.

6.1 Applicability, assumptions and limitations

The computational approach taken in this thesis allows a detailed, flexible and realistic way of modeling the hospital domain. The modeling of the hospital domain considered in this thesis comprises the complex and stochastic patient pathways and the decentralized decision-making of the different hospital units. The underlying patient pathways and decision policies can be easily adjusted which is only possible to a limited extent in the mathematical models in the existing literature on hospital patient flow logistics. The dependency of the patient flow on the available resource capacity at the different parts in the network of care units, the decision and reservation policies for specific patient types at the care units, etc. render a realistic representation of the decision-making in this domain which has not been incorporated in earlier approaches. A computational approach thus enables a realistic description of the problem and domain characteristics which facilitates better tailored solutions for patient flow logistics issues and thus promotes the applicability of the optimized solutions in hospital practice.

Furthermore, our approach enables us to consider the dynamics of patient flow logistics as opposed to the static problem settings typically addressed in the fields of operations research and operations management. Since decisions in hospital flow logistics are typically time-dependent, i.e. a decision taken now may influence the future, incorporating the problem dynamics is essential for realistically evaluating the performance of the developed techniques. Moreover, our approach allows us to consider online decision-making through which resource allocation decisions can be flexibly adjusted to better respond to changes in the environment. This adaptability and the design of the developed methods allow the proposed policies to outperform static solutions. In the design of the policies we harnessed the current hospital practice which also employs planning flexibility, which further facilitates the implementation of the proposed methods in practice.

However, there are also some critical remarks to be made. While the

model of the patient pathways and scheduling policies of the care units in Chapter 2 incorporates hospital reality to a great extent, some assumptions should be reflected upon. The probability distribution for modeling the patient LoS in our model implicates that no prior information is available when a patient transfer is to be expected. In reality the patient's clinical condition may provide an indication on the remaining LoS of the patient. In discussion with CHE domain experts, however, this possibility was excluded for the patient groups considered in the case study. Therefore, a probability distribution is deemed appropriate to model patient LoS in a general setting.

Also, while our model considers the dependency of the patient routing on the resource availability, it accounts for the dependency of the length of stay on the available resource capacity only to a limited extent. As we argued above, in practice the length of stay of a patient is affected by the patient's clinical condition, but also by the available resource capacity and the demand for care, especially at the intensive care unit. In the adaptive re-transfer mechanism, patients that have been admitted to a care unit that is not intended for by the respective pathways may be retransferred at a later point of time if the resource utilization is "high". In other respects, however, we abstracted away from this dependency based on the following considerations. In general, the criteria for initiating a patient transfer due to the need of the occupied bed may differ between hospitals and possibly even between units and care professionals within the same hospital. This supports the applicability of a probability distribution for modeling the length of stay as a realistic representation in a general hospital setting. Specifically, in the case study that was used as an instantiation for the simulation experiments conducted in this thesis, the intensive care unit had the possibility to use back-up capacity or to require additional resource capacity depending on the allocation policy to overcome resource shortage.

The coupling of patient arrival processes to the resource capacity assumed in Chapter 5, that leads to an increase in care demand for increasing resource availability, corresponds to hospital reality, but it was not possible to accurately model the coupling after the CHE case study. Deriving a model of the elasticity of patient demand for the relevant patient groups would involve extensive modeling of patient preferences, incentives offered by health insurance companies or government, the competitive (regional) landscape and the economic context of the health care market which is beyond the scope of this thesis. Potentially, our model slightly exaggerates the coupling effect to illustrate the valuable contribution of anticipation in online multi-objective decision making. For the implementation of the developed anticipatory allocation policies in a real-life hospital setting, the

modeling steps outlined above are required in order to realistically represent the respective situation.

Furthermore, the detailed modeling possibilities may require great effort for analyzing the corresponding hospital data. In a hospital setting this effort can be considerable since the quality of the data, as encountered in the case study performed in the course of this thesis, is often poor. Moreover, expert knowledge may be necessary in the modeling, e.g. for modeling the underlying routing probabilities for the patient pathways or the transfer and scheduling policies employed by the different care units. Although for applying the presented methods process modeling is only required once, a regular review of the models is advisable since changes in treatment protocols, for example, may affect the treatment processes and thus the performance of the planning techniques in practice. The publications on modeling in the health care management science area that appeared to an increasing extent in the past years may be of substantial assistance and promote automatic modeling possibilities in hospital information systems.

In our approach we assumed that allocated resources are always available and fully staffed. In hospital reality, however, this may not always be the case, for example due to illness of the staff. Then, hospital management typically advances personnel shifts from later periods of time to fill the current gap in the staff schedule, which is currently done at the case study hospital. In consequence, a gap in the available staff will arise at a later moment of time which causes a reduced resource capacity. Applying the exchange mechanism proposed in this thesis, however, would present a far better solution. According to the mechanism, capacity could be shifted to the respective unit by another unit in order to compensate the reduced resource availability in a resource neutral way if available. The exchange mechanism thus allows for implementable flexible resource adjustments and has the additional advantage of being robust to changes in the environment.

Moreover, the dynamic changes of the resource allocation induced by the techniques presented in this thesis may also require procedural changes for the hospital staff. This may require the staff to flexibly change their place of work at another unit. Although our techniques were designed to account for the present flexibility of the hospital, the implementation in practice may provide for adjustments being performed more consistently and thus imply an increased frequency of adjustments compared to current practice. In order to achieve the more frequent changes to the planning and resource allocation, the willingness of the medical staff to comply to this increased

flexibility is a precondition. Since trained personnel is scarce, this issue may be important for the actual implementation of the techniques and appropriate incentives for the staff may need to be introduced concurrently.

The resource allocation policies for flexible hospital resource management in this thesis were designed assuming frequent changes of the allocated resource capacity. In situations where a low personnel flexibility does not allow for frequent adjustments, the proposed allocation policies may need to be adjusted. One possibility would be to refine the step-function employed in this thesis using shorter utilization rate intervals and corresponding adjustments. Also, non-linear functions could be considered for determining the required resource adjustments. Another possibility would be to employ the proposed exchange mechanism as the resource adjustments are considerably less frequent than for the adaptive policies.

Another critical issue related to our approach is the computationally expensive simulation and optimization in the different settings presented in this thesis. Although the prediction using forward simulation and its application in admission control can be performed in reasonable time on a single PC, the multi-objective optimization for resource management requires considerably more computational resources which have a high-performance IT infrastructure commonly not available in hospitals. Reducing the complexity of the optimization would be necessary if the optimization needs to be performed frequently due to model changes. This issue will be discussed in further detail in the following section on possible future work.

6.2 General conclusions and possibilities for future research

In order to develop decision support techniques for hospital patient flow logistics that can be implemented in hospital practice, a realistic modeling of the domain is mandatory. Due to the complexity and dynamics of decentralized hospital organizations with stochastic treatment processes that typically involve multiple, partly shared care units, often the only option is simulation. Due to the variety of processes and the hospitals' individual ways of working, benchmark models from existing literature are typically not available to the modeler. Therefore, the simulation model should be validated to practice which often implies great effort. Moreover, using a realistic model also entails the need for solution techniques that are able

to handle the complexity and uncertainty. Our research results support the usefulness of the computational methods presented in this thesis as an effective means for improving current hospital planning and providing flexible decision support for patient flow logistics. Especially the flexibility to respond to varying situations in practice is an important aspect to be taken into account when designing implementable techniques. Furthermore, we would like to emphasize the importance to consider the network character of hospital organizations and not focus on single units. Since treatment processes typically involve multiple units, the optimization of single units potentially results in a deteriorated situation for other units and possibly the system as a whole.

Extending the findings reported in this thesis, some interesting possibilities for future research arise which will be outlined below.

For practical purposes the runtime of the multi-objective optimization in Chapter 4 and Chapter 5 should be further reduced in order to be able to perform the optimization in a more commonly available IT infrastructure in a shorter time. Ideally, the optimization should be performed within a day on a single PC enabling the re-optimization of the allocation policies if changes in the hospital environment appear to significantly change the performance of the optimized policies. Through our analysis concerning the required population size and number of allowed generations in the EDA, a decrease in runtime of the optimization of at least 75% could be achieved. A further reduction of the runtime of which the largest proportion is needed for evaluating a solution using (forward) simulation could be reduced through employing learning techniques such as neural networks to approximate the outcomes of the simulation.

Also the robustness of the multi-objective optimization solutions is an interesting area for future research. Since the objective functions considered in this thesis are stochastic and depend on the realization of the patient flows, the proposed approach should be extended by taking also the variability of the performance measures into account in the optimization. One way to achieve this might be to include the standard deviation of the performance measures as additional objectives to be minimized in the optimization. However, this would increase the complexity of the optimization problem and lead to an increased runtime of the multi-objective optimization. Another approach might be to take the the objective functions' probability distributions into account in the optimization. For this aim, multi-dimensional confidence intervals could be of use to be incorporated in the optimization. However, as these have not been defined and researched so far in the exist-

ing literature, future work is needed in order to develop a proper means of performing stochastic multi-objective optimization.

The joint optimization of patient admission control and resource allocation is also an interesting extension of the work presented in this thesis. We presented a first approach for dynamic admission control that is based on a fixed resource allocation and aims at maximizing the patient flow while considering the available capacity. An interesting question is how to combine policies for adaptive admission control and resource management to align the admission and allocation decisions. This means that patient admissions would anticipate on the future availability of resources while the resource allocation policy would take the planned future patient admissions into account for the allocation decision. Additionally, incorporating staff planning to match the staffing needs to the resource allocation could further enhance decision support and could provide interesting possibilities for future research.

Another interesting extension would be to apply the computational methods with the policy-based allocation optimization approach to other resource allocation problems. Here, potential application areas are characterized by heterogeneous and stochastic resource usage which involves multiple resources. One possibility would therefore be to apply the techniques to patient logistics in other health care settings, e.g. outpatient clinic settings with heterogeneous multi-resource treatment processes.

The computational approach presented in this thesis was ultimately designed for decision support in real-life hospital settings. Therefore, it would be interesting to perform a pilot study at a Dutch hospital. In the pilot-study, the requirements for implementing and interlinking the simulation, prediction methods and adaptive decision policies with existing hospital information systems and further automation should be assessed. It would be interesting to investigate how online feedback from reality could be incorporated in the developed methods, for example the supervised learning techniques for predicting future resource occupancy. Furthermore, the issue could be addressed how the hospital decision maker should choose among the set of Pareto optimal solutions obtained by applying the resource allocation techniques proposed in this thesis. Possibly some user-specific constraints or preferences, e.g. concerning the patient mix, could help define regions of interest in the Pareto front to reduce the magnitude of possible solutions and facilitate the final selection. Also, the patient mix that is achieved by the different allocations could be an important decision criterion. Moreover, it should be evaluated how useful practitioners consider the proposed methods to be in practice. In a preliminary study, the needs for flexibility of the

staff were evaluated with regard to the personnel's willingness to work at different units and other organizational constraints. The study showed that additional hospital-specific staffing constraints may need to be considered when shifting resources between units which limits the organization's flexibility to respond to fluctuations in supply and demand for hospital care. Although the exchange mechanism presented in this thesis accommodates with this issue through less frequent resource adjustments, it would be interesting to assess and incorporate further staffing requirements in the adaptive allocation approach.

Appendix A

Tabulated numerical results

A.1 Prediction of hospital resource usage

In this section the numerical results of the supervised learning approach presented in Chapter 3 are presented.

Unit, k	**0.7**	**0.75**	**0.8**	**0.85**	**0.9**	**0.95**
CTS-HC $k=1$	0.8302	0.6317	0.77	0.9651	0.1278	1.9008
CTS-HC $k=2$	0.7079	0.6317	0.6722	0.5643	0.6635	0.8659
CTS-HC $k=3$	0.6484	0.5897	0.7214	0.6071	0.696	0.9214
CTS-HC $k=4$	0.65	0.5825	0.7222	0.6127	0.7246	0.9611
IC $k=1$	0.1836	0.2774	0.3426	0.5748	0.7651	0.871
IC $k=2$	0.1837	0.2787	0.3221	0.3779	0.4345	0.3226
IC $k=3$	0.1446	0.2767	0.3795	0.3948	0.3945	0.3324
IC $k=4$	0.126	0.2379	0.3795	0.3948	0.4129	0.385
IC-HC $k=1$	0.4881	0.381	0.1294	0.146	0.5579	1.0389
IC-HC $k=2$	0.6151	0.5746	0.4238	0.423	0.5452	0.946
IC-HC $k=3$	0.6333	0.5841	0.4405	0.454	0.646	0.9397
IC-HC $k=4$	0.6437	0.5992	0.4484	0.4492	0.827	0.946
MC $k=1$	0.6873	0.6817	0.7333	0.696	0.6508	1.0278
MC $k=2$	0.6484	0.5357	0.5341	0.5683	0.2881	0.7071
MC $k=3$	0.6929	0.577	0.6492	0.6643	0.3373	0.7325
MC $k=4$	0.6833	0.5913	0.6968	0.6825	0.3373	0.7802
CTS-ward $k=1$	2.1579	2.6921	3.2714	4.1035	4.95	6.3667
CTS-ward $k=2$	0.477	0.4937	0.4563	0.4651	0.4995	0.5635
CTS-ward $k=3$	0.5008	0.5278	0.5087	0.5214	0.5325	0.5625
CTS-ward $k=4$	0.504	0.55	0.5283	0.546	0.5659	0.5524

Table A.1: $MAE(k,\hat{\theta})$ of measured and estimated q-quantile values for the different units for unconstrained admission control

Unit, k	**0.7**	**0.75**	**0.8**	**0.85**	**0.9**	**0.95**
CTS-HC $k = 1$	0.924	0.8166	0.973	1.0581	0.9344	1.515
CTS-HC $k = 2$	0.7985	0.746	0.7713	0.7592	0.7645	0.9359
CTS-HC $k = 3$	0.7431	0.7179	0.877	0.8522	0.915	1.0252
CTS-HC $k = 4$	0.855	0.8125	0.958	0.8861	0.977	1.1466
IC $k = 1$	0.4263	0.4307	0.495	0.5436	0.5913	0.6348
IC $k = 2$	0.2893	0.2855	0.2959	0.3169	0.3232	0.3218
IC $k = 3$	0.2798	0.3091	0.3491	0.3588	0.351	0.3251
IC $k = 4$	0.2755	0.2944	0.3236	0.3477	0.3502	0.332
IC-HC $k = 1$	0.4185	0.3976	0.2542	0.1964	0.4982	0.894
IC-HC $k = 2$	0.631	0.6601	0.5881	0.5667	0.6185	0.9173
IC-HC $k = 3$	0.647	0.6839	0.6131	0.6012	0.6958	0.9173
IC-HC $k = 4$	0.6536	0.6869	0.6137	0.6036	0.8131	0.9363
MC $k = 1$	0.2417	0.2398	0.2439	0.2678	0.3113	0.368
MC $k = 2$	0.2796	0.3037	0.334	0.3699	0.3737	0.3757
MC $k = 3$	0.2657	0.2962	0.3424	0.3865	0.377	0.3874
MC $k = 4$	0.2693	0.3	0.3461	0.3841	0.369	0.3787
CTS-ward $k = 1$	1.431	1.7732	2.1411	2.6125	3.2601	4.1113
CTS-ward $k = 2$	0.5250	0.4631	0.4911	0.5173	0.553	0.5911
CTS-ward $k = 3$	0.5208	0.4821	0.5387	0.553	0.5899	0.5738
CTS-ward $k = 4$	0.5321	0.4935	0.55	0.5827	0.6155	0.5833

Table A.2: $MAE(k, \hat{\theta})$ of measured and estimated q-quantile values for the different units for constrained admission control

Unit	**0.7**	**0.75**	**0.8**	**0.85**	**0.9**	**0.95**
MLP						
CTS-HC	0.9619	0.7667	0.9206	0.8421	0.9452	1.7119
IC	0.9825	0.946	1.8637	1.8627	1.8294	2.5508
IC-HC	0.5032	0.6833	0.9342	0.7659	0.75	1.0405
MC	0.6865	1.0611	0.8325	0.5984	0.1984	0.9849
CTS-ward	2.0913	2.03	2.2778	2.2484	2.5016	2.7921
RBN						
CTS-HC	0.5389	0.5849	0.6286	0.7627	0.7532	0.8246
IC	0.096	0.2056	0.3468	0.4984	0.854	0.7167
IC-HC	0.6619	0.7302	0.6476	0.5762	0.6127	0.8468
MC	0.6469	0.4183	0.4437	0.5349	0.1373	0.0238
CTS-ward	1.3444	1.3437	1.3977	1.3698	1.4302	1.4714
GRNN						
CTS-HC	0.8397	0.5976	0.6571	0.7754	0.7397	0.7127
IC	0.123	0.2849	0.4754	0.4944	0.8635	0.6468
IC-HC	0.5548	0.6365	0.6357	0.5698	0.6167	0.8913
MC	0.6349	0.4151	0.2659	0.4016	0.1373	0.033
CTS-ward	0.7643	0.7579	0.8048	0.85	0.8675	0.9627
Allocation-based benchmark heuristic						
CTS-HC	1.4294	0.9333	0.9905	1.1540	1.2357	1.2357
IC	4.0444	3.0992	3.1421	2.1492	2.1873	1.4508
IC-HC	1.0508	0.654	0.9349	0.9849	1.0413	1.0492
MC	0.7103	0.4183	0.2635	0.4706	0.9103	1.0238
CTS-ward	7.5595	5.9445	5.35	4.0	2.9944	2.4794

Table A.3: $MSE(\tilde{\theta})$ of measured q-quantile values and output of the different trained ANNs for unconstrained admission control averaged over 10-fold crossvalidation for basic scenario

Unit	**0.7**	**0.75**	**0.8**	**0.85**	**0.9**	**0.95**
MLP						
CTS-HC	1.3401	1.1069	1.3466	1.3150	1.3830	2.2510
IC	1.6667	1.6	1.5339	1.7435	1.678	1.5637
IC-HC	1.028	1.012	1.022	1.094	1.5083	1.6375
MC	1.2795	1.246	1.2321	1.1933	1.0746	1.7317
CTS-ward	4.4119	4.3476	4.7714	4.7238	5.0833	5.5012
RBN						
CTS-HC	1.1628	1.1522	1.1668	1.1911	1.2332	1.3231
IC	2.0702	2.1012	2.1304	2.1423	2.0988	2.0673
IC-HC	0.872	0.9548	1.1077	1.1774	1.2405	1.2571
MC	0.5911	0.6174	0.6406	0.646	0.6254	0.6839
CTS-ward	3.547	3.6173	3.6899	3.7363	3.7482	3.8078
GRNN						
CTS-HC	0.7911	0.7976	0.7976	0.7935	0.7126	0.7158
IC	0.5274	0.497	0.553	0.5736	0.5976	0.5548
IC-HC	0.6304	0.6565	0.5792	0.5869	0.7494	1.0702
MC	0.4188	0.4638	0.5045	0.5375	0.5339	0.5348
CTS-ward	0.6917	0.6673	0.7018	0.6935	0.7708	0.8381
Allocation-based benchmark heuristic						
CTS-HC	1.4494	1.1765	1.1773	1.2016	1.2462	1.2502
IC	2.55	2.6185	2.5565	2.697	2.5357	2.6423
IC-HC	0.9196	0.9048	1.0577	1.0946	1.1542	1.1649
MC	0.5871	1.2259	1.1964	1.1272	0.9875	0.8879
CTS-ward	4.1946	5.0333	5.4988	6.7786	8.2744	9.6625

Table A.4: $MSE(\tilde{\theta})$ of measured q-quantile values and output of the different trained ANNs for constrained admission control averaged over 10-fold crossvalidation for basic scenario

Unit	**0.7**	**0.75**	**0.8**	**0.85**	**0.9**	**0.95**
MLP						
CTS-HC	1.6475	1.2673	1.2028	1.7685	1.7252	2.7143
IC	1.4820	1.5087	2.0693	2.1720	2.2233	2.7628
IC-HC	0.7690	1.0526	1.0098	1.0725	1.1545	1.7651
MC	0.6865	1.0611	0.8325	0.5984	0.1984	0.9849
CTS-ward	5.3982	5.2185	5.6587	5.4915	5.9027	6.2755
RBN						
CTS-HC	0.6593	0.6978	0.7573	0.8090	0.7758	0.7463
IC	0.1890	0.2635	0.4090	0.5488	0.8658	0.7955
IC-HC	0.5926	0.6048	0.6500	0.6323	0.6971	0.9497
MC	0.6469	0.4183	0.4437	0.5349	0.1373	0.0238
CTS-ward	4.1690	4.2150	4.2943	4.4313	4.6598	4.8233
GRNN						
CTS-HC	0.7895	0.7345	0.8130	0.9022	0.7758	0.7463
IC	0.1712	0.2775	0.4122	0.4890	0.8065	0.7483
IC-HC	0.5685	0.5765	0.6354	0.5762	0.6071	0.9148
MC	0.6349	0.4151	0.2659	0.4016	0.1373	0.033
CTS-ward	0.8255	0.8320	0.8190	0.8715	0.9105	1.0060
Allocation-based benchmark heuristic						
CTS-HC	1.3825	1.0940	1.1830	1.1895	1.2438	1.2443
IC	3.6262	3.4895	2.7050	2.1220	1.7640	1.4213
IC-HC	0.8704	0.8895	1.0511	0.9865	1.0101	1.0567
MC	0.7103	0.4183	0.2635	0.4706	0.9103	1.0238
CTS-ward	7.6892	6.4115	5.8425	5.2928	4.8935	4.8233

Table A.5: $MSE(\tilde{\theta})$ of measured q-quantile values and output of the different trained ANNs for unconstrained admission control averaged over 10-fold crossvalidation for sensitivity analysis

Unit	**0.7**	**0.75**	**0.8**	**0.85**	**0.9**	**0.95**
MLP						
CTS-HC	1.8510	1.5020	1.4945	2.2290	2.2238	3.1978
IC	1.8634	1.8543	1.8714	1.8822	2.1349	2.1134
IC-HC	0.9004	1.3093	1.2278	1.2438	1.4077	2.0502
MC	1.3788	1.3517	1.3363	1.2680	1.8095	1.7668
CTS-ward	4.6840	5.1228	5.0103	5.4613	5.9035	6.2632
RBN						
CTS-HC	1.4355	1.4088	1.4185	1.4293	1.4920	1.5920
IC	2.5862	2.6110	2.6240	2.6309	2.6713	2.6292
IC-HC	0.9990	1.0227	1.0307	1.0635	1.2532	1.4563
MC	0.7243	0.7425	0.7585	0.7413	0.7370	0.7390
CTS-ward	6.5132	6.5318	6.4865	6.1525	6.2917	6.4733
GRNN						
CTS-HC	0.7519	0.7504	0.7543	0.7931	0.6880	0.7578
IC	0.8858	0.8874	0.9423	0.9790	1.0174	0.9859
IC-HC	0.6003	0.6578	0.5854	0.5525	0.8218	1.0723
MC	0.5058	0.5220	0.5368	0.5498	0.5698	0.5628
CTS-ward	0.7335	0.7418	0.7413	0.7717	0.8162	0.8427
Allocation-based benchmark heuristic						
CTS-HC	1.5043	1.2255	1.2502	1.2530	1.2818	1.2840
IC	3.1950	3.1313	2.6679	2.4841	2.1528	2.2210
IC-HC	0.9108	0.9029	1.0279	1.1716	1.0055	1.2065
MC	0.7858	1.1248	1.0872	1.1710	1.0473	0.9055
CTS-ward	5.6738	5.2247	5.1383	5.3555	5.6928	6.1680

Table A.6: $MSE(\tilde{\theta})$ of measured q-quantile values and output of the different trained ANNs for constrained admission control averaged over 10-fold crossvalidation for sensitivity analysis

Bibliography

[1] I. Adan, J. Bekkers, N. Dellaert, J. Vissers, and X. Yu. Patient mix optimisation and stochastic resource requirements: A case study in cardiothoracic surgery planning. *Health Care Management Science*, 12(2):129–141, 2009.

[2] U. Aickelin and K.A. Dowsland. An indirect genetic algorithm for a nurse-scheduling problem. *Computers & Operations Research*, 31(5): 761–778, 2004.

[3] H. Beaulieu, J.A. Ferland, B. Gendron, and P. Michelon. A mathematical programming approach for scheduling physicians in the emergency room. *Health Care Management Science*, 3(3):193–200, 2000.

[4] M. Becker and H. Czap. Artificial software agents as representatives of their human principals in operating-room-team-forming. In *Multiagent Engineering Theory and Applications in Enterprises*, pages 221–237. Springer, 2006.

[5] M. Becker, K. Krempels, M. Navarro, and A. Panchenko. Agent–Based Scheduling of Operating Theaters. In *Proceedings of the EU–LAT E–Health-Workshop*, 2003.

[6] J.W.M. Bertrand, J.C. Wortmann, and J. Wijngaard. *Production Control: A Structural and Design Oriented Approach.* Elsevier Science Inc., 1990.

[7] C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

[8] J.T. Blake and M.W. Carter. Surgical process scheduling: a structured review. *Journal of the Society for Health Systems*, 5(3):17–30, 1997.

[9] J.T. Blake and M.W. Carter. A goal programming approach to strategic resource allocation in acute care hospitals. *European Journal of Operational Research*, 140(3):541–561, 2002.

[10] B. Bollobàs. *Graph Theory: An Introductory Course*. Springer-Verlag, 1979.

[11] P.A.N. Bosman. *Design and Application of Iterated Density-Estimation Evolutionary Algorithms*. PhD thesis, Utrecht University, the Netherlands, 2003.

[12] P.A.N. Bosman and J. Grahl. Matching inductive search bias and problem structure in continuous estimation-of-distribution algorithms. *European Journal of Operational Research*, 185(3):1246–1264, March 2008.

[13] P.A.N. Bosman and J.A. La Poutré. Learning and anticipation in online dynamic optimization with evolutionary algorithms: the stochastic case. In *GECCO '07: Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 1165–1172, New York, NY, USA, 2007. ACM.

[14] P.A.N. Bosman and D. Thierens. Multi–objective optimization with diversity preserving mixture–based iterated density estimation evolutionary algorithms. *International Journal of Approximate Reasoning*, 31:259–289, 2002.

[15] P.A.N. Bosman and D. Thierens. Adaptive variance scaling in continuous multi-objective estimation-of-distribution algorithms. In *GECCO '07: Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 500–507, New York, NY, USA, 2007. ACM.

[16] P.A.N. Bosman, J. Grahl, and F. Rothlauf. SDR: a better trigger for adaptive variance scaling in normal edas. In *GECCO '07: Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 492–499, New York, NY, USA, 2007. ACM.

[17] P.A.N. Bosman, J. Grahl, and D. Thierens. Enhancing the performance of maximum-likelihood gaussian edas using anticipated mean shift. In *Parallel Problem Solving from Nature - PPSN X*, pages 133–143, 2008.

[18] J. Bowers and G. Mould. Managing uncertainty in orthopaedic trauma theatres. *European Journal of Operational Research*, 154(3):599–608, 2004.

[19] L. Braubach, W. Lamersdorf, Z. Milosevic, and A. Pokahr. Policy-rich multi-agent support for e-health applications. In *Challenges of Expanding Internet: E-Commerce, E-Business, and E-Government*, pages 235–249. Springer, 2005.

[20] L.T. Bui, H.A. Abbass, and J. Branke. Multiobjective optimization for dynamic environments. In *Congress on Evolutionary Computation*, volume 3, pages 2349–2356, 2005.

[21] E. Burke, P. De Causmaecker, G. Vanden Berghe, and H. Van Landeghem. The state of the art of nurse rostering. *Journal of Scheduling*, 7:441–499, 2004.

[22] M.W. Carter and S.D. Lapierre. Scheduling emergency room physicians. *Health Care Management Science*, 4(4):347–360, 2001.

[23] J.R. Charnetski. Scheduling operating room surgical procedures with early and late completion penalty costs. *Journal of Operations Management*, 5(1):91–102, 1984.

[24] H. Czap and M. Becker. Multi-agent systems and microeconomic theory: A negotiation approach to solve scheduling problems in high dynamic environments. volume 3, page 83b, Los Alamitos, CA, USA, 2003. IEEE Computer Society.

[25] K. Deb. *Multi objective optimization using evolutionary algorithms*. Wiley, 2001.

[26] K. Decker and J. Li. Coordinated hospital patient scheduling. In *Proceedings of the Third International Conference on Multi-Agent Systems (ICMAS98)*, pages 104–111, 1998.

[27] K.S. Decker. *Environment Centered Analysis and Design of Coordination Mechanisms*. PhD thesis, University of Massachusetts, 1995.

[28] K.S. Decker and V.R. Lesser. Designing a family of coordination algorithms. In *Proceedings of the First International Conference on Multi-Agent Systems*, pages 73–80, 1995.

[29] S.A. DeLoach, M.F. Wood, and C.H. Sparkman. Multiagent Systems Engineering. *The International Journal of Software Engineering and Knowledge Engineering*, 11(3):231–258, 2001.

[30] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[31] A Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27:642–669, 1956.

[32] M. Farina, K. Deb, and P. Amato. Dynamic multiobjective optimization problems: test cases, approximations, and applications. *Evolutionary Computation, IEEE Transactions on*, 8:425–442, 2004.

[33] R.B. Fetter and J.L. Freeman. Diagnosis related groups: Product line management within hospitals. *The Academy of Management Review*, 11(1):41–54, 1986.

[34] P.A. Fishwick. *Simulation model design and execution : building digital worlds*. Prentice Hall, 1995.

[35] F. Gorunescu, S.I. McClean, and P.H. Millard. Using a queueing model to help plan bed allocation in a department of geriatric medicine. *Health Care Management Science*, 5(4):307–312, 2002.

[36] M. Greenwood. The natural duration of cancer. *Reports on Public Helath and Medical Subjects*, 33:1–26, 1926.

[37] P.M.A. Groot. *Decision Support for Admission Planning under Multiple Resource Constraints*. PhD thesis, Eindhoven University of Technology, the Netherlands, 1993.

[38] N. Guinet and S. Chaabane. Operating theatre planning. *International Journal of Production Economics*, 85(1):69–81, 2003.

[39] E. Hans, G. Wullink, M. van Houdenhoven, and G. Kazemier. Robust surgery loading. *European Journal of Operational Research*, 185(3): 1038–1050, 2008.

[40] P.R. Harper and A.K. Shahani. Modelling for the planning and management of bed capacities in hospitals. *Journal of the Operational Research Society*, 53:11–18, 2002.

[41] G.W. Harrison. Implications of mixed exponential occupancy distributions and patient flow models for health care planning. *Health Care Management Science*, 4(1):37–45, 2005.

[42] R. Herrler and F. Klügl. Simulation. In *Multiagent Engineering Theory and Applications in Enterprises*, pages 575–596. Springer, 2006.

[43] R. Herrler and F. Puppe. Adaptivity and scheduling. In S. Kirn, P. Herzog, O. Lockemann, and O. Spaniol, editors, *Multiagent Engineering Theory and Applications in Enterprises*, pages 277–299. Springer, 2006.

[44] C.-J. Ho and H.-S. Lau. Evaluating the impact of operating conditions on the performance of appointment scheduling rules in service systems. *European Journal of Operations Research*, 112:542 – 553, 1999.

[45] J.L. Hodges and E.L. Lehmann. *Basic Concepts Of Probability And Statistics*. Society for Industrial and Applied Mathematics, 2nd edition, 2004.

[46] M. Hunink and P. Glasziou. *Decision making in health and medicine: Integrating evidence and values*. Cambridge University Press, 2001.

[47] A.K. Hutzschenreuter, P.A.N Bosman, I. Blonk-Altena, J. van Aarle, and J.A. La Poutré. Agent-based patient admission scheduling in hospitals. In Nishiyama Berger, Burg, editor, *Proc. of 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008) - Industry and Applications Track*, pages 45–52, 2008.

[48] A.K. Hutzschenreuter, P.A.N. Bosman, and J.A. Poutré. Evolutionary multiobjective optimization for dynamic hospital resource management. In *EMO '09: Proceedings of the 5th International Conference on Evolutionary Multi-Criterion Optimization*, volume 5467 of *Lecture Notes in Computer Science*, pages 320–334. Springer-Verlag, 2009.

[49] A.K. Hutzschenreuter, P.A.N. Bosman, and J.A. La Poutré. Enhanced hospital resource management using anticipatory policies in online dynamic multi-objective optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference – GECCO 2010*, to appear.

[50] G.R. Iversen and H. Norpoth. *Analysis of variance*. Sage Publications, Inc, 1986.

[51] S.D. Izenberg, M.D. Williams, and A. Luterman. Prediction of trauma mortality using a neural network. *American Surgeon*, 63:275–281, 1997.

[52] G. Kaandorp and G. Koole. Optimal outpatient appointment scheduling. *Health Care Management Science*, 10(3):217–229, 2007.

[53] E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[54] E. Kaplansky and A. Meisels. Distributed personnel scheduling - negotiation among scheduling agents. *Annals of Operations Research*, 2005.

[55] S. Kim, I. Horowitz, K.K. Young, and T.A. Buckley. Flexible bed allocation and performance in the intensive care unit. *Journal of Operations Management*, 18(4):427–443, 2000.

[56] S. Kirn, C. Anhalt, H. Krcmar, and A. Schweiger. Agent.hospital – health care applications of intelligent agents. In S. Kirn, P. Herzog, O. Lockemann, and O. Spaniol, editors, *Multiagent Engineering Theory and Applications in Enterprises*, pages 199–220. Springer, 2006.

[57] K.J. Klassen and T.R. Rohleder. Scheduling outpatient appointments in a dynamic environment. *Journal of Operations Management*, 14:83 – 101, 1996.

[58] P. Kolesar. A markovian model for hospital admission scheduling. *Management Science*, 16(6):384–396, 1970.

[59] A. Kumar, P.S. Ow, and M.J. Prietula. Organizational simulation and information systems design: an operations level example. *Management Science*, 39(2):218–240, 1993.

[60] R.J. Kusters and P.M.A. Groot. Modelling resource availability in general hospitals design and implementation of a decision support model. *European Journal of Operational Research*, 88(3):428–445, 1996.

[61] S. Littig and M. Isken. Short term hospital occupancy prediction. *Health Care Management Science*, 10:47–66, 2007.

[62] W.E. Lowell and G.E. Davis. Predicting length of stay for psychiatric diagnosis-related groups using neural networks. *Journal of the American Medical Informatics Association*, 1:459–466, 1994.

[63] A. Marazzi, F. Paccaud, C. Ruffieux, and C. Beguin. Fitting the distributions of length of stay by parametric models. *Medical Care*, 36 (6):915–927, 1998.

[64] C.C. Marinagi, C.D. Spyropoulos, C. Papatheodorou, and S. Kokkotos. Continual planning and scheduling for managing patient tests in hospital laboratories. *Artificial Intelligence in Medicine*, 20(2):139–154, 2000.

[65] A. Marshall, C. Vasilakis, and E. El-Darzi. Length of stay-based patient flow models: Recent developments and future directions. *Health Care Management Science*, 8(3):213–220, 2005.

[66] L. Maruster, T. Weijters, G. de Vries, A. van den Bosch, and W. Daelemans. Logistic-based patient grouping for multi-disciplinary treatment. *Artificial Intelligence in Medicine*, 26:87–107, 2002.

[67] A. Moreno, D. Isern, and D. Sanchez. Provision of agent-based health care services. *AI Communications*, 16(3):167–178, 2003.

[68] M. Moz and M. Vaz Pato. A genetic algorithm approach to a nurse rerostering problem. *Computers & Operations Research*, 34(3):667–691, 2007.

[69] J.L. Nealon and A. Moreno. *Agent-Based Applications in Health Care*, pages 3–18. Birkhueser Verlag, 2003.

[70] M. Nikraz, G. Caire, and P.A. Bahri. A methodology for the analysis and design of multi-agent systems using JADE. *International Journal of Computer Systems Science and Engineering*, 21(2), 2006.

[71] A. Oddi and A. Cesta. Toward interactive scheduling systems for managing medical resources. *Artificial Intelligence in Medicine*, 20 (2):113–138, 2000.

[72] D. Parr and J. Thompson. Solving the multi-objective nurse scheduling problem with a weighted cost function. *Annals of Operations Research*, 155:279–288, 2007.

[73] J. Patrick and M.L. Puterman. Improving resource utilization for diagnostic services through flexible inpatient scheduling: A method for improving resource utilization. *J Oper Res Soc*, 58(2):235–245, 2006.

[74] J. Patrick, M.L. Puterman, and M. Queyranne. Dynamic multipriority patient scheduling for a diagnostic resource. *Operations Research*, 56 (6):1507–1525, 2008.

[75] T.O. Paulussen, N.R. Jennings, K.S. Decker, and A. Heinzl. Distributed patient scheduling in hospitals. In *Proceedings of the Eigtheenth International Joint Conference on Artificial Intelligence (IJCAI-03)*, 2003.

[76] T.O. Paulussen, A. Zöller, A. Heinzl, A. Pokahr, L. Braubach, and W. Lamersdorf. Dynamic patient scheduling in hospitals. In M. Bichler et al., editors, *Coordination and Agent Technology in Value Networks*, pages 255–275. GITO Berlin, 2004.

[77] T.O. Paulussen, A. Zöller, F. Rothlauf, A. Heinzl, L. Braubach, A. Pokahr, and W. Lamersdorf. Agent-based patient scheduling in hospitals. In S. Kirn, P. Herzog, O. Lockemann, and O. Spaniol, editors, *Multiagent Engineering Theory and Applications in Enterprises*, pages 255–275. Springer, 2006.

[78] M. Pidd. *Computer simulation in management science.* Wiley, West Sussex, 1998.

[79] V. Podgorelec and P. Kokol. Genetic Algorithm Based System for Patient Scheduling in Highly Constrained Situations. *Journal of Medical Systems*, 21(6):417–427, 1997.

[80] J.C. Ridge, S.K. Jones, M.S. Nielsen, and A.K. Shahani. Capacity planning for intensive care units. *European Journal of Operational Research*, 105(2):346–355, 1998.

[81] A.V. Roth and R. Van Dierdonck. Hospital resource planning: concepts, feasibility, and framework. *Production and Operations Management*, 4(1):2–29, 1995.

[82] S.J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach.* Prentice-Hall, Inc., 1995.

[83] R. Sibbel and C. Urban. Agent-based modeling and simulation for hospital management. In *Cooperative Agents, Applications in the Social Sciences*, pages 183–202. Dordrecht, Boston, London, 2001.

[84] D. Sier, P. Tobin, and C. McGurk. Scheduling surgical procedures. *Journal of the Operational Research Society*, 48(9):884–891, 1997.

[85] E.A. Silver, D.F. Pyke, and R. Peterson. *Inventory management and production planning and scheduling.* New York: Wiley, 1998.

[86] B.W. Silverman. *Density estimation for statistics and data analysis.* Chapman & Hall, 1992.

[87] V.L. Smith-Daniels, S.B. Schweikhart, and D.E. Smith-Daniels. Capacity management in health care services: Review and future. *Decision Sciences*, 19:898–919, 1988.

[88] D.P. Strum, L.G. Vargas, J.H. May, and G. Bashein. Surgical suite utilization and capacity planning: A minimal cost analysis model. *Journal of Medical Systems*, 21(5):309–322, 1997.

[89] C. Stummer, K. Doerner, A. Focke, and K. Heidenberger. Determining location and size of medical departments in a hospital network: A multiobjective decision support approach. *Health Care Management Science*, 7(1):63–71, 2004.

[90] C.D. Sypyropoulos. AI planning and scheduling in the medical hospital environment. *Artificial Intelligence in Medicine*, 20:101–111, 2000.

[91] D. Tandberg and C. Qualls. Time series forecasts of emergency department patient volume, length of stay, and acuity. *Annals of Emergency Medicine*, 23(2):299–306, 1994.

[92] J. Thornton and A. Sattar. An integer programming-based nurse rostering system. *Concurrency and Parallelism, Programming, Networking, and Security*, pages 357–358, 1996.

[93] J. VanBerkel, P.and Blake. A comprehensive simulation for wait time reduction and capacity planning applied in general surgery. *Health Care Management Science*, 10(4):373–385, 2007.

[94] I.B. Vermeulen, S.M. Bohte, D.J.A. Somefun, and J.A. La Poutré. Improving Patient Schedules by Multi-agent Pareto Appointment Exchanging. In *Proceedings of 2006 IEEE International Conference on E-Commerce Technology (CEC/EEE 2006)*, pages 185–196, 2006.

[95] I.B. Vermeulen, S.M. Bohte, S.G. Elkhuizen, H. Lameris, P.J.M. Bakker, and J.A. La Poutré. Adaptive resource allocation for efficient patient scheduling. *Artificial Intelligence in Medicine*, 46(1):67–80, 2009.

[96] J. Vissers and R. Beech, editors. *Health Operations Management: Patient flow logistics in health care*. Health Management Series. Routledge, 2005.

[97] J.M.H. Vissers. Patient flow-based allocation of inpatient resources: A case study. *European Journal of Operational Research*, 105(2):356–370, 1998.

[98] J.M.H. Vissers, J.W.M. Bertrand, and G. De Vries. A framework for production control in health care organizations. *Production Planning and Control*, 12(14):591–604, 2001.

[99] J.M.H. Vissers, I.J.B.F. Adan, and J.A. Bekkers. Patient mix optimization in tactical cardiothoracic surgery planning: a case study. *IMA Journal of Management Mathematics*, 16(3):281–304, 2005.

[100] S. Walczak, W.E. Pofahl, and R.J. Scorpio. A decision support tool for allocating hospital bed resources and determining required acuity of care. *Decision Support Systems*, 34:445–456, 2003.

[101] G. Weiss, editor. *Multiagent systems. A modern approach to distributed artificial intelligence*. The MIT Press, 1999.

[102] E.-K. Yeong, T.-C. Hsiao, H. K. Chiang, and C.-W. Lin. Prediction of burn healing time using artificial neural networks and reflectance spectrometer. *Burns*, 31(4):415–420, 2005.

Summary

A Computational Approach to Patient Flow Logistics in Hospitals

Scheduling decisions in hospitals are often taken in a decentralized way. This means that different specialized hospital units decide autonomously on e.g. patient admissions and schedules of shared resources. Decision support in such a setting requires methods and techniques that are different from the majority of existing literature in which centralized models are assumed. The design and analysis of such methods and techniques is the focus of this thesis. Specifically, we develop computational models to provide dynamic decision support for hospital resource management, the prediction of future resource occupancy and the application thereof.

Hospital resource management targets the efficient deployment of resources like operating rooms and beds. Allocating resources to hospital units is a major managerial issue as the relationship between resources, utilization and patient flow of different patient groups is complex. The issues are further complicated by the fact that patient arrivals are dynamic and treatment processes are stochastic.

Our approach to providing decision support combines techniques from multi-agent systems and computational intelligence (CI). This combination of techniques allows to properly consider the dynamics of the problem while reflecting the distributed decision making practice in hospitals. Multi-agent techniques are used to model multiple hospital care units and their decision policies, multiple patient groups with stochastic treatment processes and uncertain resource availability due to overlapping patient treatment processes. The agent-based model closely resembles the real-world situation. Optimization and learning techniques from CI allow for designing and evaluating improved (adaptive) decision policies for the agent-based model, which can then be implemented easily in hospital practice.

In order to gain insight into the functioning of this complex and dynamic problem setting, we developed an agent-based model for the hospital care units with their patients. To assess the applicability of this agent-based model, we developed an extensive simulation. Several experiments demonstrate the functionality of the simulation and show that it is an accurate representation of the real world. The simulation is used to study decision support in resource management and patient admission control.

To further improve the quality of decision support, we study the prediction of future hospital resource usage. Using prediction, the future impact of taking a certain decision can be taken into account. In the problem setting at hand for instance, predicting the resource utilization resulting from an admission decision is important to prevent future bottlenecks that may cause the blocking of patient flow and increase patient waiting times. The methods we investigate for the task of prediction are forward simulation and supervised learning using neural networks. In an extensive analysis we study the underlying probability distributions of resource occupancy and investigate, by stochastic techniques, how to obtain accurate and precise prediction outcomes.

To optimize resource allocation decisions we consider multiple criteria that are important in the hospital problem setting. We use three conflicting objectives in the optimization: maximal patient throughput, minimal resource costs and minimal usage of back-up capacity. All criteria can be taken into account by finding decision policies that have the best trade-off between the criteria. We derived various decision policies that partly allow for adaptive resource allocations. The design of the policies allows the policies to be easily understandable for hospital experts. Moreover, we present a bed exchange mechanism that enables a realistic implementation of these adaptive policies in practice. In our optimization approach, the parameters of the different decision policies are determined using a multiobjective evolutionary algorithm (MOEA). Specifically, the MOEA optimizes the output of the simulation (i.e. the three optimization criteria) as a function of the policy parameters. Our results on resource management show that the benchmark allocations obtained from a case study are considerably improved by the optimized decision policies. Furthermore, our results show that using adaptive policies can lead to better results and that further improvements may be obtained by integrating prediction into a decision policy.

Samenvatting

Een Computationele Aanpak voor Patiëntenlogistiek in Ziekenhuizen

In ziekenhuizen worden planningsbeslissingen vaak op een decentrale manier genomen. Dit wil zeggen dat verschillende gespecialiseerde ziekenhuisafdelingen autonoom beslissen over, bijvoorbeeld, de opname van patiënten en roosters voor resources waar meerdere afdelingen gebruik van maken. Beslissingsondersteuning voor ziekenhuisplanning vereist daarom methoden en technieken die, anders dan in de bestaande literatuur, geen centrale modellen veronderstellen. In dit proefschrift richten we ons op het ontwerp en de analyse van zulke methoden en technieken. Specifiek ontwikkelen we computermodellen voor dynamische beslissingsondersteuning met betrekking tot het capaciteitsmanagement, het voorspellen van toekomstig beddengebruik en de toepassing hiervan in de besluitvorming.

Capaciteitsmanagement heeft als doel een efficiënte inzet van capaciteiten, zoals operatiekamers en ziekenhuisbedden. De toewijzing van capaciteiten aan ziekenhuisafdelingen is een belangrijk en moeilijk probleem voor het bestuur van een ziekenhuis omdat capaciteiten, hun benutting en de betreffende patiëntstromen complex met elkaar samenhangen. Bovendien wordt het probleem nog verder gecompliceerd door dynamische aankomsten van patiënten en stochastische behandelprocessen.

In onze aanpak combineren we technieken uit multi-agent systemen en computationele intelligentie (CI). Met behulp van deze combinatie van technieken kunnen we rekening houden met zowel de dynamiek als de decentrale manier van besluitvorming in ziekenhuizen. We gebruiken multi-agent technieken voor het modelleren van meerdere ziekenhuisafdelingen en hun beslissingsstrategieën, meerdere patiëntgroepen met stochastische en deels overlappende behandelprocessen. De stochasticiteit introduceert onzeker-

heden in de beschikbaarheid van capaciteiten. Dit agent-gebaseerde model is een realistische beschrijving van de werkelijkheid. Met behulp van CI technieken voor optimalisatie en leren kunnen betere (adaptieve) beslissingsstrategiën ontworpen en geëvalueerd worden voor het agent-gebaseerde model, die makkelijk in de praktijk geïmplementeerd kunnen worden.

Om een beter inzicht in dit complexe en dynamische probleem te krijgen, hebben we een agent-gebaseerd model voor de ziekenhuisafdelingen en hun patiënten ontwikkeld. Op basis van dit model hebben we een uitgebreide simulatie ontwikkeld om de toepasselijkheid van het agent-gebaseerde model te beoordelen. De functionaliteit en praktische relevantie worden gedemonstreerd met behulp van verschillende simulatie experimenten. Daarnaast wordt de simulatie gebruikt om beslissingsondersteuning op het gebied van capaciteitsmanagement en opnameplanning te onderzoeken.

We bestuderen ook het voorspellen van de toekomstige bedbezetting om de kwaliteit van de beslissingsondersteuning verder te verbeteren. Met behulp van predictie kan rekening worden gehouden met het effect van een bepaalde beslissing op de toekomst. Voor dit probleem is het voorspellen van de bedbezetting belangrijk om toekomstige capaciteitsknelpunten te voorkomen die opstoppingen van de patiëntstromen en langere wachtrijen tot gevolg kunnen hebben. In onze aanpak evalueren we twee voorspellingstechnieken: voorwaartse simulatie en leren onder toezicht met behulp van neurale netwerken. In een uitgebreide analyse bestuderen we de onderliggende kansverdelingen van de bedbezetting en onderzoeken we met behulp van stochastische technieken hoe nauwkeurige en preciese voorspellingen gedaan kunnen worden.

Bij beslissingen over capaciteitsallocatie in ziekenhuizen moet vaak met meerdere doelen rekening gehouden worden. In de optimalisatie aanpak beschouwen we drie doelen: maximale patiëntendoorstroom, minimale capaciteitskosten en het minimale gebruik van reserve capaciteit. Deze doelen worden gelijktijdig geoptimaliseerd door het vinden van beslissingsstrategieën die de verschillende doelen op vergelijkbare wijze tegen elkaar afwegen. We presenteren meerdere beslissingsstrategieën die voor een deel adaptieve allocatie beslissingen faciliteren. De strategieën zijn ontworpen met het oog op de begrijpbaarheid door medische specialisten. Bovendien introduceren we een bedden schuifmechanisme waarmee de adaptieve strategieën realistisch in de praktijk geïmplementeerd kunnen worden. Voor de optimalisatie van meerdere doelen worden de parameters van de beslissingsstrategieën met behulp van een multi-doel evolutionair algoritme (*multi-objective evolutionary algorithm* – MOEA) bepaald. Specifiek optimaliseert het MOEA de simulatieuitkomsten (dwz. de drie doelen) als functie van de

strategie parameters. Onze resultaten tonen aan dat standaardallocaties, die in de praktijk gebruikt worden, aanzienlijk verbeterd kunnen worden door de geoptimaliseerde beslissingsstrategieën. Bovendien geven de experimenten aan dat adaptieve allocatie strategieën en het gebruik van predictie informatie binnen de strategieën tot verdere verbeteringen leiden.

Acknowledgements

It is a pleasure to thank the many people who have supported me during the past years.

First of all, I want to thank my supervisors Han La Poutré, Will Bertrand and Peter Bosman. I thank Han La Poutré for giving me this opportunity, the challenging discussions and his valuable advice and input throughout the PhD project. I am grateful to Will Bertrand for his useful comments, his continuing interest in my project and his motivating feedback. I would also like to thank my co-promotor Peter Bosman for the insightful discussions and his helpful advice and input. Furthermore, I am grateful to my graduation supervisor, Ger Koole, from the VU University Amsterdam for his support and recommendation for the PhD position.

The cooperation with the Catharina Hospital in Eindhoven has been an invaluable gain in the project. It provided insights into hospital practice and allowed me to relate my research to a real-life problem in this fascinating domain. I want to give my thanks to Jan van Aarle for his helpful feedback, Ilona Blonk-Altena for her help and support, Dick Koning and Floor Haak for their comments and help, Erik Korsten for establishing the contact with the Catharina hospital and the nurses and planners for their expert knowledge and time.

A big thank you also goes to my (former) colleagues at the IS group for the pleasant past years. In particular, my thanks go to my roommate Florian Gottschalk for making the atmosphere in our office so friendly, helpful, supportive and "gezellig"; to Ana Karla Alves de Medeiros, Mariska Netjes, Irene Vanderfeesten, Maja Pešić and Ting Wang for our dinners and other activities; to Monique Jansen-Vullers for her listening ear and advice; to Paul Grefen for his support and for enabling me to frequently visit the CWI in Amsterdam; to the secretaries Ada Rijnberg, Annemarie van der Aa and Ineke Withagen for their support; moreover, I want to thank Anne Rozinat, Christian Günther, Jeroen van Luin, Sven Till, Samuil Angelov, Hajo Reijers, Jochem Vonk, Jos Trienekens, Peter van den Brand, Ronny

Mans, Jana Samalikova, Ricardo Seguel and Boudewijn van Dongen. Also, I want to thank Geertje Kramer and Annemie van der Werf from Beta for their help.

Working with my (former) colleagues from SEN4 at the CWI in Amsterdam has been very helpful and valuable. Special thanks to Han Noot for his programming help and our cordial conversations; Sander Bohte for his help and advice on neural networks; Valentin Robu, Mengxiao Wu, Sara Ramezani and Ivan Vermeulen for their friendship, our discussions, the nice coffee breaks and our "Han's4Oio" dinners. Moreover, I would like to thank the CWI and the SARA/NWO-NCF for enabling me to use the high-performance computing systems for the costly optimization computations.

As a balance to work I greatly enjoyed doing sports. I appreciate the excellent facilities provided by the TUe sport center where I could let off some steam once in a while. My thanks to the ESAC for the enjoyable rock climbing activities, especially to Mariska, Robert, Stijn, Thijs, Dirk, Koen and Fred. In addition to rock climbing, I also greatly enjoyed horseback riding during which I forgot all about the PhD. Many thanks go to Kamie for the great instructions. Also, I thank Janneke, Eefje, Jolanda, Kamille, Fabiënne, Robert and the Thursday & Friday evening groups from the Eindhovense Manege for the fun lessons and our chats afterwards. Especially, I want to thank Lipton & his friends for the great time and the fun and exciting horseback outings in the summer.

I am also grateful to my "paranymfen" Mariska Netjes and Mengxiao Wu. Dear Mariska, dear Mengxiao, thank you for accepting to support me during the defense of my dissertation. I greatly value your friendship and support throughout my PhD project in Eindhoven and Amsterdam.

Lastly, and most importantly, I wish to thank my friends and family. Thanks Alina, Anne, Stefanie, Maja and Mariska for our talks and for keeping my mind off the PhD! Thank you Marijke for your great help with the design of the cover. My late father greatly encouraged me to go abroad and to start a PhD project. He has exemplified to me to achieve my goals through perseverance, effort and hard work and, most importantly, to never look back. To him I dedicate this thesis. Liebe Mama, lieber Jens, ich bin Euch unsagbar dankbar für Eure Unterstützung, Aufmunterung und Euren unerschütterlichen Glauben in mich. Ohne Euch hätte ich die letzten Jahre nicht geschafft und ich weiss, dass wir drei immer auf einander bauen können. Liebster Christian, vielen Dank, dass Du mich all die Jahre darin unterstützt hast meinen Weg zu gehen und mir durch alle Höhen und Tiefen zur Seite gestanden bist.

Curriculum vitae

Anke Hutzschenreuter was born in Ulm, Germany, on May 1, 1980. In 1999 she started her studies in "Wirtschaftsmathematik" (business mathematics) at Ulm University, Germany. During her studies, Anke also completed a Master program on Business Mathematics and Informatics at the Vrije Universiteit Amsterdam, The Netherlands. Anke graduated in 2004 from the Vrije Universiteit Amsterdam, on the subject of simulation-based outpatient appointment scheduling under the supervision of Professor Ger Koole. In 2005, Anke graduated from Ulm University, Germany, on the subject of queueing models for outpatient appointment scheduling under the supervision of Professor Ger Koole and Professor Ulrich Rieder. In 2005, Anke started as Ph.D. student at the Eindhoven University of Technology (TU/e), The Netherlands. The research, that resulted in this thesis, was undertaken within the project 'AgI-CARE: Agent-based Models for Adaptive Intelligent Systems in Health Care Planning' under the supervision of Professor Han La Poutré. The project was performed in cooperation with the Catharina Hospital Eindhoven, The Netherlands, and the Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands.