Katholische Universität
Eichstätt-Ingolstadt

Bed Planning: Advanced Applications of Operational Research in Large Hospitals

DOCTORAL THESIS

Unterlagen gemäß § 5 Abs. 3 der Fachpromotionsordnung der Wirtschaftswissenschaftlichen Fakultät der Katholischen Universität Eichstätt-Ingolstadt sowie gemäß § 7 Abs. 3 der Rahmenpromotionsordnung der Katholischen Universität Eichstätt-Ingolstadt

Erstprüfer:	Prof. Dr. Alexander Hübner
Zweitprüfer:	Prof. Dr. Pirmin Fontaine
Eingereicht von:	Manuel Walther
Datum Einreichung:	10 April 2020
Datum Verteidigung:	27 Juli 2020

Contents

Acknowledgements

During my time as an external research assistant at the department of Supply Chain Management and Operations at the Catholic University of Eichstätt-Ingolstadt I was accompanied by various supporters to whom I would like to express my deepest gratitude.

First and foremost, I would like to thank my supervisor and co-author of three papers included in this dissertation, Prof. Dr. Alexander Hübner, who has expertly guided me during several joint research projects and has become a very good friend and mentor. Together with Prof. Dr. Heinrich Kuhn he created a very open, inspiring, and pleasant working environment. I am very thankful for the many engaging and fruitful discussions that we had together that greatly improved my research and kept me motivated throughout my work. In addition, I am grateful for having been given the opportunity as an external research assistant to work with and teach students in different OR-related subjects as well as the opportunity to supervise several Bachelor's and Master's theses. In addition, I would like to extend my special thanks to Prof. Dr. Pirmin Fontaine for co-supervising this dissertation.

A terrific research group at Prof. Dr. Hübner's and Prof. Dr. Kuhn's departments has further been a great support and helpful resource throughout my time in Ingolstadt: Birgit Jürgens, Mareike Müller, Tobias Düsterhöft, Prof. Dr. Andreas Holzapfel, Sandro Kühn, Marcel Lehmann, Tobias Potoczki, Dr. Kai Schaal, Dr. Manuel Ostermeier, Dr. Dominik Wörner, Dr. Michael Sternbeck, Dr. Andreas Popp, and Dr. Johannes Wollenburg.

I would further like to extend a very special thank you to Fabian Schäfer together with whom I have co-authored two papers included in this thesis. Fabian was wonderful to work with and has always been a great friend.

Most of all and finally, I would like to thank my wife and kids, as well as my parents for their limitless support, not only during my dissertation, but during my whole life. Without them, this research project could not have been realized.

Munich, 10 April 2020

Manuel Walther

List of Figures

List of Tables

1 Introduction

In many countries today, a rising life expectancy and the associated demographic shift, coupled with the advancements of modern medicine, has fueled an ever-increasing cost pressure on healthcare systems. A driving factor for these rising costs can be seen in inpatient stays in hospitals that in many cases are connected to cost-intensive treatments. A central concern of any hospital management in such an environment is therefore to understand how to make the best possible use of available resources. A decisive factor in this regard is the management of bed capacities.

The present cumulative dissertation comprises four contributions, which address open research questions in the field of strategic, tactical and operative bed planning:

1 Walther, M., 2020. Strategical, tactical, and operational aspects of bed planning problems in hospital environments. Submission planned to Social Science Research Network (SSRN)
2 Hübner, A., Kuhn, H., Walther, M., 2018. Combining clinical departments and wards in maximum-care hospitals. OR Spectrum 40, 679-709
3 Schäfer, F., Walther, M., Hübner, A., Kuhn, H., 2019. Operational patient-bed assignment problem in large hospital settings including overflow and uncertainty management. Flexible Services and Manufacturing Journal 31, 1012–1041
4 Schäfer, F., Walther, M., Hübner, A., Grimm, D., 2020. Machine learning and pilot method: tackling uncertainty in the operational patient-bed assignment problem. Submitted to OR Spectrum on 13 February 2020

The first contribution sets out to provide an overview over the different hierarchical planning levels on which bed planning problems may be addressed. It should be noted in this context that several different aspects may be combined under the collective term "bed planning". These may be delimited in terms of their scope and their planning horizon. A frequently used taxonomy in this context is the hierarchical subdivision of typical problems in health care into strategical, tactical and operational levels as provided by Hulshof et al. (2012). In the context of bed planning, a typical strategical problem is how to combine departments and wards to obtain benefits from pooled ward capacity. On a tactical level, an exemplary problem setting related to bed planning can be seen in devising master surgery schedules that optimize downstream bed occupancy levels as patients returning from surgery will require a bed for post-surgical recovery and monitoring. Finally, on an operational level, patient-bed allocations need to be optimized while taking the objectives and constraints of patients and medical staff alike into account.

To start, the second contribution deals with the strategical problem of combining departments into groups and assigning pooled ward capacity to these groups with the goal of balancing bed occupancy levels within a hospital. Specifically, one of the underlying goals is to minimize the amount of beds required to meet a predetermined service level. However, merging ward capacities with the aim of simultaneously accommodating patients from different medical departments increases the complexity of organizing and ensuring proper care for these patients. This leads to so-called pooling costs. To tackle this problem, a modeling and solution approach is developed which is based on a generalized partitioning problem and is solved by integer linear programming (ILP). This enables hospital management to determine the cost-optimal combination of all departments and wards in a hospital, while ensuring that predetermined thresholds with regard to maximum walking distances for doctors and patients are adhered to.

Once pooled ward capacities are established, the solution space for allocating incoming patients to beds is greatly increased and the underlying allocation

problem quickly becomes too complex to be handled without computational support. In this regard the third contribution ties in with the second contribution in that it deals with optimizing the operational patient-bed allocation problem. In order to enable optimal allocation of patients to beds, it is important to identify and take into account the individual needs and limitations of the three main stakeholders involved, namely patients, doctors, and nursing staff. All of these stakeholders exhibit different and sometimes contradicting objectives and constraints, such that a trade-off has to be made that maximizes the overall utility for the hospital. In addition, the complexity of the problem is increased by the high volatility and uncertainty regarding patient arrivals, types of illnesses, and the resulting remaining lengths of stay of newly arriving patients. In order to address this situation, a mathematical model and solution approach for the patient-bed allocation problem is developed that is designed to generate solutions for large, real-life operative planning situations. In addition to being able to deal with overflow situations, this solution approach further takes different patient types into account, for example by anticipating emergency patient arrivals.

Finally, the fourth contribution builds on the third contribution in that the modeling and solution approach to allocate patients to beds is extended by several aspects. As mentioned above, hospitals have to deal with uncertainty regarding the actual demand for beds. Here, the fourth contribution improves the anticipation of emergency patients by using machine learning. Specifically, weather data, seasons, important local and regional events, and current and historical occupancy rates are combined to better anticipate emergency inpatient arrivals. In addition, a hyper-heuristic approach is developed based on the pilot method defined by Voß et al. (2005). By combining the improved anticipation of emergency patients with this hyper-heuristic approach significant improvements can be achieved compared to the solution approach presented in the third contribution.

2 Strategical, Tactical, and Operational Aspects of Bed Planning Problems in Hospital Environments

Submission planned to *SSRN*

Abstract With rising health care costs, fueled by an aging demographic as well as pharmaceutical and technical advances in medicine, hospitals have to ensure that all available resources are put to use as efficiently as possible. A key aspect in this regard is the management of bed capacities. This includes strategical, tactical, and operational planning problems that are hierarchically dependent on one-another and that are based on different time frames, namely years, months, and days, respectively. This paper exemplifies and introduces a selected number of these problems across the different hierarchy levels to provide an introduction to the subject. On a strategical level overall hospital capacity and layout planning are discussed. With regard to tactical problem settings, the impact of adapting master surgery schedules on bed occupancy levels is highlighted. This is then complemented by an overview of two operational problems, namely the scheduling of elective inpatients and the allocation of patients to beds upon arrival.

2.1 Introduction

In modern medicine, hospital resources have to be put to use in the most efficient way possible as healthcare costs are steadily rising due to an aging demographic as well as advances in modern medicine which lead to costlier treatments. In terms of bed management in hospitals, there are several aspects that need to be considered, such as the need to increase overall occupancy levels for bed capacities while ensuring a predetermined availability, or how to efficiently distribute incoming patients to rooms and beds. This paper is designed to give an idea of the variety of different bed planning aspects and highlights key issues we have encountered during our work on bed-planning related operational research questions and during several joint research projects with a large German hospital. It is designed to link the most relevant practical insights with literature, but not to detail all aspects of hospital bed management.

To better understand the different stakeholders and aspects to be considered in the broader context of bed planning in a hospital, one may look at the typical patient path in a hospital which is schematically exemplified in Figure 2.1. To begin with, a typical large hospital has two types of patients that arrive on any given day, namely elective patients and emergency patients. Elective patient arrivals can be planned whereas emergency arrival rates are by definition unpredictable. Both of these patient types may either require ambulant care, meaning that they can have a consultation with a physician or receive treatment on the same day they arrive, or they may require longer term monitoring that requires at least one overnight stay. Patients that leave on the same day as they arrive are typically called outpatients, whereas patients that stay over night are referred to as inpatients. Although outpatients may sometimes require a bed for certain procedures, the actual bed capacity of any hospital typically refers to their overnight bed capacity. This includes all rooms and their respective beds that are located in dedicated inpatient wards. It should be noted in this

context, that when discussing bed planning problems, this paper only refers to inpatient bed capacity and not to the actual physical beds themselves.

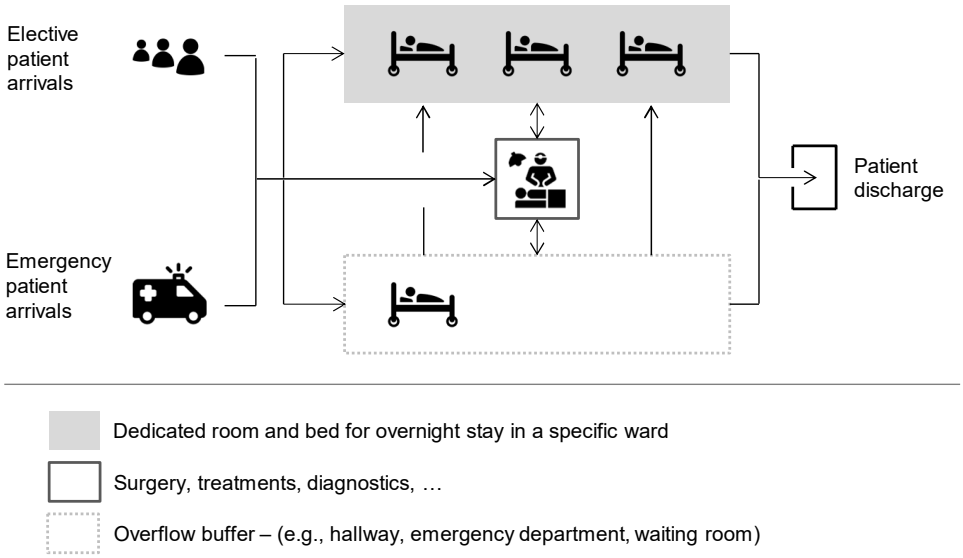


Figure 2.1: Schematic overview of an inpatient path through a hospital

Coming back to Figure 2.1, it should be noted that "bed planning" should be seen as a collective term that touches on many different planning problems other than merely allocating patients to beds. For instance, some inpatients will undergo surgical procedures and/or diagnostic treatments during their hospital stay. After such treatments, they will be required to remain in the hospital for recovery or follow-up procedures. Thus, scheduling surgical and diagnostic treatments directly affects bed capacity utilization. Furthermore, overall bed capacity in a hospital needs to be adequately sized to provide sufficient ward space to meet the downstream bed capacity needs of such surgical and diagnostic treatment facilities in order to minimize the occurrence of overflow situations in which patients have to temporarily stay in overflow buffer areas such as hallways or waiting rooms.

The above-mentioned aspects merely provide examples for what needs to be considered when talking about holistic bed management approaches. In essence, it doesn't seem feasible to design a "one-size fits all" bed management approach. Instead, different models can be formulated for

different problem settings. In this regard, some problems are hierarchically dependent on one another. Hulshof et al. (2012) have designed a taxonomic classification to categorize different planning decisions in health care. They differentiate between strategical, tactical, and operational problem settings. These problem settings differ in terms of planning hierarchy and time frame. Specifically, strategical, tactical, and operational problems as well as the implications of their potential solutions, are considered in terms of years, months, weeks and days, respectively. With regard to bed planning and management, Table 2.1 highlights typical problem settings along said hierarchical categories that will further be elaborated on in the following sections of this paper.

planning hierarchy	time horizon	typical problem settings (exemplary)
strategical	years	▪ planning/building overall hospital capacity and infrastructure ▪ pooling resources in department/ward-combinations
tactical	months	▪ adapting master surgery schedules to level bed occupancy rates
operational	weeks	▪ scheduling elective patients to level out occupancy over time
	days	▪ allocating rooms/beds to patients

Table 2.1: Examples for hierarchical classification of bed planning problems - planning hierarchy based on Hulshof et al. (2012)

On a strategic level, bed planning considerations play an important role when considering hospital layouts. For instance, the hospital layout as a whole should be designed such that walking distances for patients and staff alike are minimized. Specifically, wards should be designed and allocated to departments in a way that enables patients to be as close as possible to their respective required treatment and diagnostic facilities. Likewise, the same considerations can be made for doctors having to tend to patients in wards and having to perform certain treatment procedures. For example, the catheterization lab should be placed in close vicinity to the cardiology

wards. In addition, pooling bed capacities across departments may help balance bed occupancy levels.

On a tactical level, bed occupancy levels are directly affected by upstream planning and scheduling decisions for surgical and diagnostic procedures. For example, when scheduling patients for an elective surgery, the typical length of stay (LOS) post-surgery needs to be considered. However, operating room schedules also depend on staff rosters, cleaning procedures and the like, such that a trade-off has to be determined between optimizing downstream bed capacity levels and achieving the most efficient use of surgical space.

On an operational level, different types of patients have to be considered who have different requirements regarding available bed capacity. To give an example, depending on the severity and type of illness a specific patient exhibits, certain infrastructural requirements may have to be met. This might range from being medically isolated due to a contagious disease to having a bed which is equipped with telemetry infrastructure for patients with certain cardio-vascular diseases. Another example can be seen in that elective patients will typically not understand why a bed is not available to them on the day of their scheduled arrival, whereas emergency patients will be more likely to understand the need to be held in an overflow area until a respective room and bed opens up. Furthermore, any hospital can only be run with dedicated and highly trained staff. In this context it is important that the respective objectives and constraints of doctors, nurses, technicians and other staff are adequately met. For instance, this might entail grouping certain types of patients in a specific ward area to ensure efficient rounds by minimizing walking distances for doctors.

The remaining paper is structured as follows. Section 2.2 deals with strategical aspects in hospital bed management. Sections 2.3 and 2.4 discuss tactical and operational problems in bed planning while highlighting the impact of surgery scheduling on bed management as well as operational patient-bed

allocation, respectively. Finally, section 2.5 provides a brief summary and discusses potential future developments in bed management.

2.2 Strategical aspects of bed management

When talking about strategical aspects of bed management in large hospitals, the term "strategical" is used to refer to the highest planning hierarchy as defined by Hulshof et al. (2012). Although all bed planning aspects involve some type of "strategy" this section focuses solely on decision problems related to bed planning and management for which the time horizon is to be seen in years (see Table 2.1). Specifically, the following sub-sections discuss general aspects of hospital layout planning with regard to bed management as well as the main benefits and challenges of pooling bed capacities.

2.2.1 Hospital layout planning with regard to bed management

When designing a hospital a potential starting point is to anticipate the amounts and types of patients that will be treated within the facilities. For large maximum-care hospitals this may include 20 and more different medical specialties or dedicated departments (see for example Hübner et al. (2018)) including for example pediatrics, obstetrics, orthopedics, urology, neurology, vascular surgery and so forth. All of these departments cater to different types of patient clientele and have different infrastructural requirements.

After having been admitted to a specific department and ward, patients are prepped for a specific surgery, diagnostic procedure, or medical treatment. Depending on the type of procedure that a specific patient has to go through, a certain time for recovery and/or further treatment has to be planned.

In terms of bed capacity required, one can take these data points and determine the amount of beds required for a certain number of operating theatres, diagnostic machines (MRIs, CTs, etc.) and the like.

Finally, after determining the amount of beds required to satisfy the individual needs of each of the medical departments, the question of locating bed capacities within a hospital remains. In this context, two key aspects have to be considered. First, patients as well as doctors have to move within the hospital. Patients have to be transported to and from the OR while doctors have to do rounds, i.e., tend to their patients, and – depending on their specialty – may have to move back and forth between the emergency rooms, the operating theatres and the wards. Helber et al. (2015), for example, investigate ways to optimize these walking distances using a hierarchical layout planning approach to minimize logistic costs which are directly linked to patient transportation.

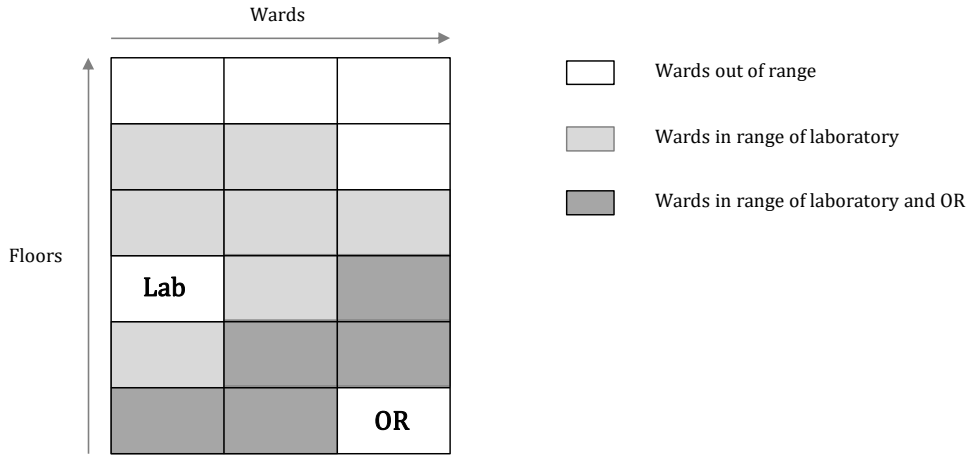


Figure 2.2: Schematic overview of maximum allowable distances in a hospital

Second, some wards have to be in close vicinity to certain diagnostic facilities or the OR. To give an example, the obstetrics ward should be fairly close to the delivery rooms. Depending on what facilities are required for individual wards and patients, a hospital planner can thus design the hospital by optimizing walking distances while at the same time ensuring that maximum distance thresholds are not surpassed (see for example

Hübner et al. (2018)). To illustrate this, Figure 2.2 shows a schematic overview of a exemplary hospital comprising six floors with three wards each. With the configuration shown, only five wards could potentially be used for catering to patients that require being in the vicinity of the lab and the ORs.

2.2.2 Pooling bed capacities

One of the key challenges many hospitals face is how to efficiently make use of their bed capacities. Oftentimes, specific wards will fill up on weekdays while other wards might have free capacities left. In this context, simply relying on the total utilization of the overall hospital bed capacity may be misleading as this does not account for the amount of overflow situations occurring locally within specific departments and wards. A better indicator for efficient use can thus be seen in the combined evaluation of bed availability, i.e., the probability of a department not having enough dedicated ward space for a given patient, and overall bed capacity utilization.

Faced with this challenge, many hospitals nowadays are considering pooling bed capacities to reap the benefits of balancing effects. To give an example, Van Essen et al. (2015) and Hübner et al. (2018) have investigated methods to combine ward space and assign the resulting combined space to groups of departments. Here, the rationale is to investigate the median distribution of occupied beds per department over time while finding departments that can be matched. An example for this can be seen in Figure 2.3. Here, three different departments are considered which each exhibit different occupancy distributions during a typical week. Although the average number of beds occupied per department is equal to 30 in all three cases, the actual number of beds required to meet the needs of a combined department is quite different. Specifically, when combining departments 1 and 2, 74 beds will be required to meet the required average bed capacity for this combination

on a typical Wednesday. However, when combining departments 1 and 3, this number can be reduced to 65 for this example.

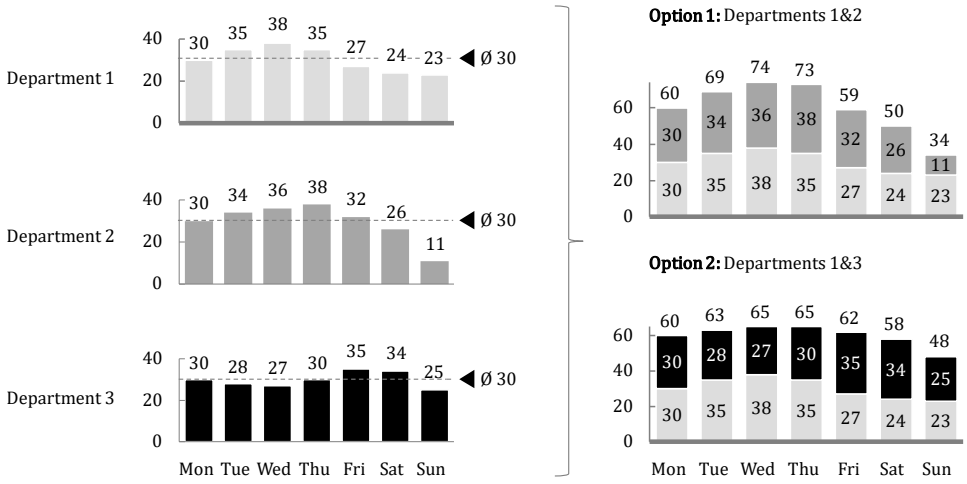


Figure 2.3: Combining clinical departments to balance occupancy levels according to Hübner et al. (2018)

To reap the benefits of combining ward capacity for specific groups of departments, it is however imperative that the patient clientele of the departments in question are actually compatible with each other as this means assigning patients of different departments to the same ward or room. This compatibility covers several aspects. First, the patient clientele itself has to be compatible. For example, it would not be possible to combine pediatrics with general surgery, as children are typically separated from adults in a hospital. Likewise, the obstetrics department would not be combined with other departments as women in labor and newborn babies require their own space. Second, combined ward space also requires the attending nursing staff to be trained accordingly so that they can cater to all patients from both departments. Furthermore, all respective rooms and beds have to be equipped with the relevant infrastructure, e.g., telemetry which is needed for the respective department combination in question. To put it in a nutshell, there is a trade-off between the different levels of grouping departments and wards which needs to be assessed. This relationship is illustrated by Figure 2.4.

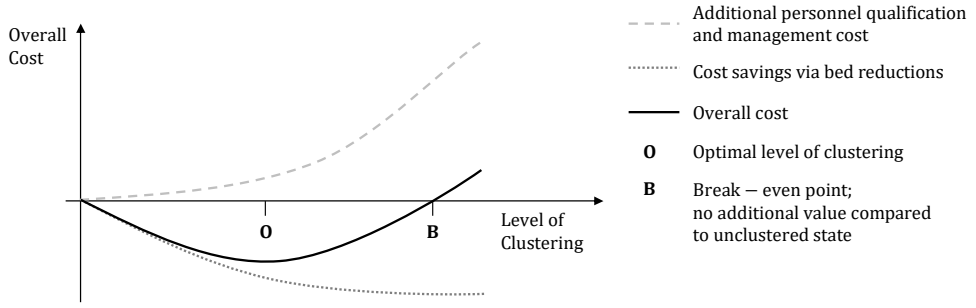


Figure 2.4: Schematic cost-curve when combining departments and wards to pool bed capacities

In essence, combining departments and assigning pooled ward space to said group of departments entails additional costs for qualifying nursing staff as well as providing additional infrastructure. On the other side of the equation, pooling ward capacity reduces the number of beds, staff and other resources that have to be held available to meet a certain availability target assuming a given utilization rate. There are however limits to the benefits of pooling bed capacities. As can be seen in Figure 2.4, the overall cost actually rises compared to a non-combined state if one were to assume a maximum amount of pooled beds. This is because combining departments and wards makes sense for beneficial combinations in which the "clustering cost", i.e., additional qualification and infrastructure expenses, are low while at the same time combining departments exhibiting complementary occupancy level distributions. However, when creating the largest theoretically possible combination of departments, one would also combine unfavorable combinations that would create more costs than benefits.

2.3 Tactical aspects of bed management

Having a well-thought out hospital layout in which overall bed capacities, operating room capacities, available infrastructure and so forth are designed to meet the needs of the local patient clientele based on long-term analysis of historical distributions is only one part of the story. As pointed out in previous sections, it is imperative for hospital management to not only ensure high overall occupancy rates over time with regard to their bed capacity but to make sure that the availability of said bed capacity stays high as well. This requires balancing of bed occupancies over time and across wards. Balancing bed occupancies across wards is a strategical problem related to combining departments and assigning pooled ward capacity to these combinations which has been discussed in previous section 2.2. When it comes to balancing bed occupancies over time, there are several aspects involved. In theory, one could simply anticipate the total LOS for each elective patient based on their planned procedure, respectively. This data could then be used to rearrange the arrivals of elective patients in such a way that it further takes anticipated emergency arrivals in the near future as well as current occupancies into account, such that bed occupancies are leveled over time. In addition, elective patient arrivals could theoretically be rescheduled on an ad-hoc basis depending on the current occupancy situation.

In practice, the operating rooms (ORs) and certain diagnostic infrastructure also need to be used as efficiently as possible, as the cost to keep such infrastructure up and running is substantial. The underlying thought process is the same as described above for bed capacities. To illustrate this aspect, one can consider a typical OR. To efficiently make use of OR capacity means to ensure a very high utilization, i.e., ensure that every OR is actually used for surgical procedures during normal operating hours, avoid delays and quickly shift around patients if necessary. Elective surgical procedures can theoretically be scheduled in a way that optimizes OR occupancy while avoiding overflow situations as average cutting-suture times are known. In

reality, however, staff constraints as well as downstream effects of patient movements have to be considered. In particular, the required surgeons and support staff have to be on duty whereas patients recovering from surgery require a bed where they can stay after surgery. Working plans for doctors and nurses are typically drawn up weeks or months in advance to help staff better plan and bring together their private and professional lives. This doesn't mean that every specific surgical procedure has to be scheduled weeks in advance, but it requires to determine time slots in which specific types of procedures, i.e., procedures that can be done by a specific team of physicians can be performed. These time slot schedules are typically called "master surgery schedules" and have to be decided on months in advance.

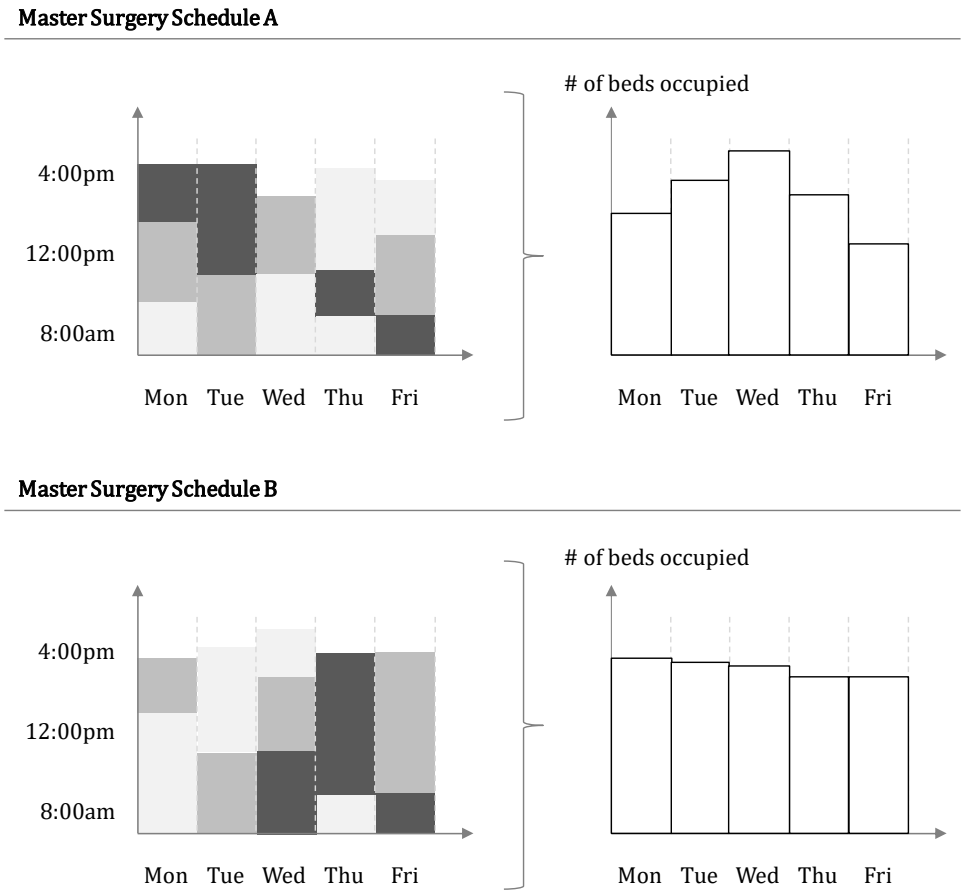


Figure 2.5: Schematic example for the effects of master surgery schedules on bed occupancies

Figure 2.5 shows a schematic example for the effects that different master surgery schedules may have on bed occupancies. Specifically, every type of surgical procedure requires different recovery and surveillance periods for the patients that have undergone surgery resulting in respective bed occupancies. In the schematic example shown in Figure 2.5, two master surgery schedules A and B are compared. Both surgery schedules show similar OR utilization rates and allow for scheduling the same amount of surgical procedures per week. Here, each shade of gray represents time slots that may be filled with specific types of surgeries for patients from specific departments, respectively. However, as can be seen in the bed occupancy graphs on the right, the resulting bed capacity requirements significantly differ between both schedules. Depending on the overall utilization of bed capacity within the hospital, it is clear that optimizing master surgery schedules has the potential to significantly impact bed capacity requirements and should therefore be considered when assessing bed management on a tactical level (see for example Demeulemeester et al. (2013) or Fügener et al. (2014)). Similar considerations apply when looking at schedules for certain diagnostic treatments such as MRI and PET.

2.4 Operational aspects of bed management

The operational aspect of bed management is what usually comes to mind when talking about bed management in a hospital, i.e., the actual allocation of patients to beds. It should be stressed, however, that this is only the last piece in the puzzle as laid out in the previous sections. In terms of planning hierarchy (see Table 2.1, it is crucial that the strategical and tactical aspects of bed planning have been thoroughly assessed and taken into account before optimizing the operational aspect of bed management. In other words, optimizing bed occupancy levels simply by changing the way elective patients are scheduled for arrival at the hospital cannot by itself make due for a lack of organization and planning on the strategical and tactical level.

On an operational level, bed management can be split into two different sub-tasks which themselves can be seen as hierarchical decisions. Namely, the first task is scheduling elective inpatients for arrival while the second task is allocating a physical room and bed to an inpatient who has just arrived.

2.4.1 Scheduling elective inpatient arrivals

The first task does not require the respective admission office of a specific medical department to assign a specific bed to the patient. Instead, quite similar to hotel room reservations, the key question at this stage is whether or not enough capacity will be available on a certain day in the future.

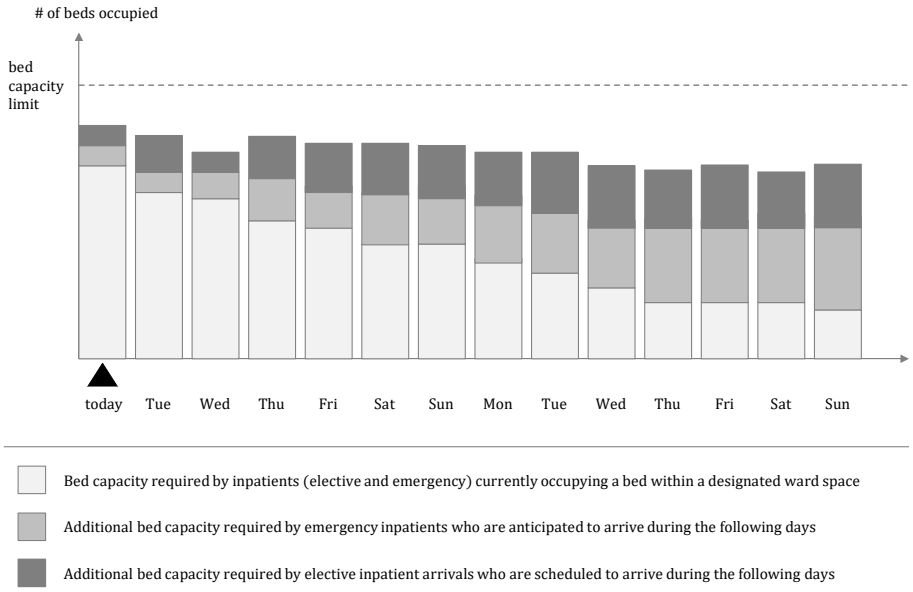


Figure 2.6: Example for bed occupancies over time when scheduling elective patients

Figure 2.6 provides a schematic graph which shows the total anticipated bed occupancies over time divided by categories. Specifically, each bar depicts the amount of beds that are required for a certain group of patients on a particular day. This data is obtained by combining the anticipated LOS per patient with the amount of patients requiring a bed. For instance,

the light gray bars shows the bed capacity required for inpatients (elective and emergency) that are already in the hospital at the time of planning, i.e., patients that already occupy a physical room and bed. These patients are scheduled to leave the hospital within the following days, such that the light gray bar gets smaller over time. In the example of Figure 2.6, two weeks after today there are only about 20% of the patients anticipated to still be in the hospital who are there today. The gray and dark gray bars on the other side depict the total amount of beds required for future inpatient arrivals, and are typically growing over time. This data is anticipated based on the average LOS for the respective emergency and elective inpatient arrivals. In the graph, one can see that the overall number of beds that at present are believed to be required for the coming two weeks is slowly declining. This is due to the fact that in this particular example there is still room for scheduling elective inpatient arrivals within the next two weeks. However, it should also be noted that the admission office in a hospital typically tries to stay under the overall bed capacity limit by a predetermined safety margin to avoid overflow, i.e., to account for statistical volatility in both emergency patient arrival rates as well as changes in LOS. Naturally, scheduling elective patients does not only require anticipating future bed occupancy levels but also involves planning surgeries and accounting for special infrastructural needs, i.e., telemetry, a specific patient might have, which for individual patients will change the solution space as to which day these patients may be scheduled to arrive.

2.4.2 Allocating patients to beds

Allocating patients to rooms and beds is an everyday task which is typically carried out by the admission office once a patient arrives at a hospital and requires a bed for an overnight stay. Revisiting the hotel room reservation analogy from the previous subsection, it should be stressed that the upstream decision of scheduling elective inpatients only takes the overall bed capacity that meets the requirements for the specific patient into account.

The actual physical allocation to a bed will then happen only at the exact moment a patient physically requires his or her respective room and bed. When allocating a patient to a bed, several stakeholders are involved, all of whom have different objectives and constraints. Namely, these are the patients themselves, the nursing staff, and doctors (see Schäfer et al. (2019)).

	Beds	Rooms	Wards
			O Objectives C Constraints
Patient	O Minimize waiting time for beds for elective and emergency inpatients C Ensure infrastructural requirements	O Maximize compatibility between patients (age, medical condition) C Prohibit gender mixing C Ensure medical isolation & infrastructural requirements	
Nursing Staff			O Balance workload for nursing staff across wards
Doctor		O Minimize walking distances for doctors during rounds	

Table 2.2: Overview of typical objectives and constraints for patients, nurses, and doctors in the operational patient-bed allocation problem

Table 2.2 gives an overview of objectives and constraints that usually have to be met in this regard according to Schäfer et al. (2019). To begin with, several hard constraints have to be adhered to when allocating inpatients to rooms and beds. First, women and men are typically not allowed to be mixed in the same room (except for certain intensive care units) for an overnight stay. Second, all medically required infrastructure needs to be available for a certain patient at a given bed. For example, certain cardiac patients "require telemetry-ready" beds such that their cardiac status can be constantly monitored by the attending nurses. Third, medical isolation requirements have to be respected. This includes, for example, isolating immuno-comprimised patients from normal patient clientele, or isolating patients that have especially contagious or dangerous diseases such as MRSA-infected patients. When all these necessary hard constraints are met, the solution space for allocating a patient to a bed can then be assessed with regard to the objectives of the three stakeholders. Patients as such will want their time spent in the hospital to be as comfortable as possible

while receiving the best treatment and care. A good proxy for ensuring comfort can be maximizing compatibility between room mates with regard to their medical condition and their age (see also Schäfer et al. (2019)). Finally, elective patients may rightfully assume that a bed is available for them on the day they are scheduled to arrive. Rescheduling elective patient arrivals on short notice is only a theoretical option, as it would cause problems for most patients. This is because such patients will usually have undergone significant planning to account for their hospital stay in their private and professional lives. Thus, forcing elective inpatients to be rescheduled on short notice will leave them unhappy with the hospital's service and drive them away to other hospitals. However, as hospitals are forced to operate with overall average bed occupancy rates of 80% and more, overflow situations will happen due to the volatility of emergency arrival rates, patient LOS and so forth. When overflow situations occur, incoming emergency patients might sometimes be redirected to other hospitals if they are still en route to the ER. Nonetheless, this does not prevent overflow situations as patients who are already in the hospital might unexpectedly have to stay longer. Another example could be seen in emergency patients who have to unexpectedly spend the night after a first diagnostic treatment has revealed the need for prolonged monitoring. When overflow situations occur, some patients may be required to be temporarily placed in overflow areas such as hallways, waiting rooms, or treatment rooms until a bed within their respective department ward space becomes available. To avoid having to reschedule elective patients or having to put them in such overflow areas, a good option is to prefer elective patients when assigning bed capacity provided that all medically required treatments may still be upheld for all patient types alike. To summarize, the key to mitigating overflow situations is to avoid them altogether and – if that is not feasible – to manage them as best as possible.

Nursing staff are typically assigned to a specific ward and cannot be moved around easily. This is because the work requires well-coordinated teams that have an in-depth understanding of the specific procedures within a certain department or ward. In addition, nursing staff cannot be quickly

reassigned to different shifts as this would contravene personal planning capabilities for the individual staff members. As a result, the care capacity for a given ward on a given day is limited on short notice, i.e., the staff assigned to a specific ward on a specific day are only capable of handling a certain care work load. In terms of bed management this translates to avoiding to assign too many care intensive patients to a single ward. As different patients have different requirements regarding daily care, the key is to balance care workload across different wards within the same pooled ward capacity in line with the respective care capacity available in each of these wards.

Finally, the allocation of patients to rooms and beds also effects doctors. Typically, doctors have a set of patients they are assigned to and whom they attend to at least once a day during rounds. The idea with regard to bed management is to reduce walking distance for doctors such that patients that are assigned to a specific doctor are located in close vicinity to one another, preferably even sharing rooms.

Taking all the above-mentioned objectives and constraints into account when planning patient allocations for pooled ward capacity of 50 beds or more will become almost impossible to do by hand, especially considering the high levels of uncertainty that are inherent to almost any patient clientele with regard to LOS, arrival times, undetected illnesses and so forth. To handle such problems, mathematical optimization models have been proposed by numerous research groups (see for example Demeester et al. (2010), Ceschia and Schaerf (2016), and Schäfer et al. (2019)). Moreover, in addition to using automated assignment algorithms, professional software tools are required that show the bed planner, nurse or doctor all the key information relevant for patient bed allocation at a glance for a respective combination of wards.

In case a fully integrated tool that automatically assigns patients to beds does not yet exist, it can still be very helpful to provide a bed planner with an

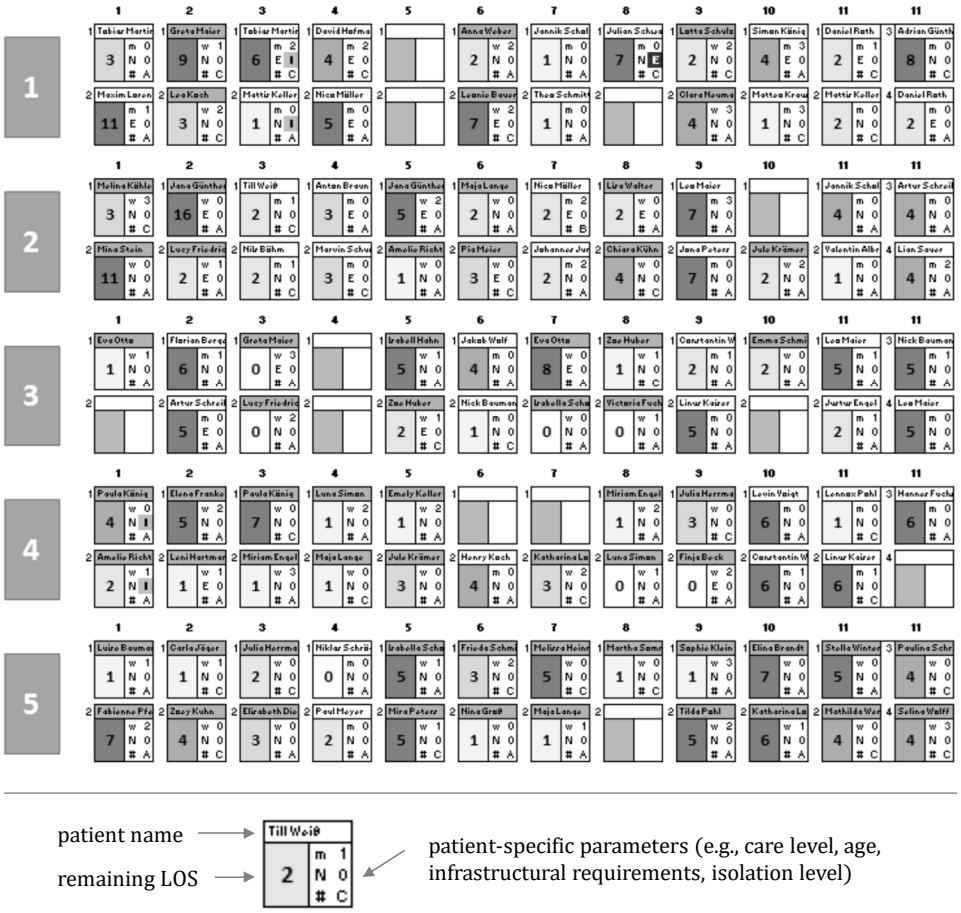


Figure 2.7: Example of an overview tool designed to support a bed manager in a hospital

overview that shows him or her all relevant information required to support patient-bed allocation decisions at a glance. For example, Figure 2.7 depicts an example view of such a basic tool, in which all occupants within 120 beds spanning across five wards are shown with their respective parameters relevant for bed planning purposes, such as age, gender, remaining LOS, infrastructural needs, or care level. In the example shown in Figure 2.7 every ward has a total of 12 double bed rooms that can be occupied by either only male or only female patients. For each bed, a name on a white background signifies a male patient, whereas a name on a gray background signifies

a female patient. In this example, only room five in ward one and room four in ward three are completely empty whereas five beds are available in rooms already having one female occupant and three beds are available in rooms already having one male occupant. Depending on the gender of known future patient arrivals, the bed planner can use this information to decide whether or not to start filling an empty room. Furthermore, the remaining LOS is depicted in the lower left corner. For example, the female occupant in bed two in room ten in ward two is anticipated to leave the hospital in two days, whereas the male occupant in bed one in room eight in ward one is anticipated to leave in seven days. With this information readily available the bed planner can combine patients with similar remaining LOS such that both beds in a given room are vacated on the same day, thereby providing more flexibility when assigning future patients to beds as an empty room can be used either for male or female patients. Finally, having an overview of patient-specific parameters, such as the required care-level, infrastructural needs, and so forth, depicted in the lower right corner, enables the bed planner to balance workload for nursing staff and provides a quick overview over patient needs.

2.5 Summary and discussion

Rising health care costs have been forcing hospitals more than ever before to operate as efficiently as possible. First and foremost this includes optimizing the allocation of hospital resources as well as anticipating workload. In a large hospital, this requires the careful orchestration of multiple medical departments, diagnostic facilities, medical and non-medical staff, and last but not least bed capacity. All of these different stakeholders have their own objectives and constraints, respectively, that are closely interlinked with each other. Naturally, not all of these objectives can be optimized at the same time due to conflicting interests of the individual stakeholders involved. Hence, trade-offs are required when looking to optimize the overall utility for the hospital.

In this context, bed management plays an important role in hospital planning activities. “Bed management” as such is a broad term encompassing several individual planning problems. These problems may be categorized along their planning hierarchy, i.e., with regard to the typical time frames in which problems occur, decisions can be made, and results achieved. In particular, these problems can be categorized in strategical, tactical, and operational aspects, wherein the time horizon for each of these can be viewed in years, months, and days, respectively. The strategical side mainly includes hospital layout aspects as well as questions with regard to pooling bed capacities across departments and wards. On a tactical level, the interlink between main cost drivers such as diagnostic facilities and operating rooms and bed capacity and bed management is assessed, while the operational level then deals with the actual ad-hoc allocation of patients to beds. However, the different problem settings on each hierarchical level cannot be combined in a single mathematical model designed to deliver solutions for all levels simultaneously. This is mainly due to the fact that the cost for adopting a solution to a specific problem setting, in terms of money and time spent to implement that solution, is substantially different for every hierarchical level. For example, once the layout of a hospital has been decided on, and after the hospital has been built according to that layout plan, an actual change of that layout will most likely require time- and cost-extensive remodeling efforts. Likewise, the decision to form combinations of wards and departments with the goal of creating pooled ward capacity has a significant lead time as qualification measures for nursing staff have to be set up and implemented, for example. In essence, this means that for each of these hierarchical layers, the underlying problem settings are only worth solving again, should the relevant input parameters change systematically. For example, it would not be sensible to regroup departments and wards in case of a one-time surge of inpatients, e.g., due to a large highway accident, whereas it would make sense to contemplate such a regrouping in case a long-term trend in patient arrivals for a specific department is observed, e.g., triggered by closures of smaller hospitals in the vicinity. However, this does not mean that the effects that a hierarchically higher decision has on downstream problem settings should

be neglected. On the contrary, it simply means that decision variables for mathematical models of problems on different hierarchical layers are unlikely to be determined within the same optimization model since the values for said variables require changing in substantially different intervals. For instance, a specific master surgery schedule typically stays the same for at least a couple of weeks while patient-bed-allocations may change within hours. Instead, the connections between different planning levels need to be formulated as boundary conditions or specific targets. For example, when optimizing said master surgery schedule it is crucial to balance average bed occupancy levels.

The key takeaway for any hospital management in this regard is to understand that there are intricate relationships between all these aforementioned aspects without making the error of trying to tackle all of the underlying planning problems in one big "one size fits all" optimization model. Instead, the key is to look at these problems hierarchically, while being aware of the general implications that certain decisions on a higher hierarchical level have on downstream planning problems. One prominent example here is the effect master surgery schedules have on bed capacity requirements.

The main challenge that hospitals will face is the question of how to put theory into practice. This is due to the rising need for computational support when running larger bed management and planning systems comprising several hundred rooms and beds. Many hospitals today are in the midst of a digital transformation, trying to move away from a paper-based organization of patient pathways and treatment plans to a fully digital system. In this transformation process it will be crucial to holistically look at the combination of planning problems with regard to bed management without focusing on specific medical departments or diagnostic facilities individually, but on the hospital as a whole. As different stakeholders will have contradicting objectives, it is key that the development and oversight of the necessary digital infrastructure and planning software is done on a hospital level, thereby preventing "fights for resources" and providing planning security for all departments involved.

3 Combining Clinical Departments and Wards in Maximum-Care Hospitals

Co-authors: Alexander Hübner, Heinrich Kuhn
Published in *OR Spectrum* on 19 May 2018

Abstract Sharing bed capacity across clinical departments improves bed availability via pooling effects. This means in effect that fewer beds are required to satisfy a given service level when combining departments and wards into groups. However, this increases the complexity of tending to inpatients and therefore creates what we term pooling costs. To solve the trade-off, we suggest an integer linear programming (ILP) modeling and solution approach that is designed on a generalized set partitioning problem (SPP). The approach finds the cost-minimal combination of departments and wards in a maximum-care hospital that satisfies maximum walking distance thresholds for doctors and patients. In particular, costs associated with holding the required bed capacity are minimized while also considering seasonality of weekly demand as well as personnel qualification costs and management costs incurred by combining departments and allocating pooled ward capacity to these combinations. In addition, maximum walking distances between wards and central facilities for the combinations obtained are minimized. As actual management practice justifies solving the conflicting criteria of the entire planning problem in lexicographical order, the resulting hierarchically structured and SPP-based ILP approach additionally allows for time-efficient and exact solutions even for large problem settings. Our modeling and solution approach was co-developed and implemented at a large German maximum-care hospital comprising 22 clinical departments. As a result, the number of beds needed to maintain a unified service level of 95% can be reduced by 3.3%, while cutting costs by 2.1%. We also perform several sensitivity analyses and show general applicability by using simulated data for generalized and very large hospital settings.

3.1 Introduction

Hospital resources have to be used as efficiently as possible while maintaining a sufficient level of patient care and optimizing the workload for medical staff. High and stable bed occupancy levels play a key role in this context. Today, most hospitals are experiencing rising caseloads paired with cuts in overall bed capacity. These tight bed capacities are required from an economic efficiency perspective, but can lead to bed shortages, which can in turn result in a cumbersome search for unoccupied beds within other departments of the hospital, long waiting times for patients, or even lead to patient rejection. As a result, most large maximum-care hospitals – which are by definition obliged to treat any incoming patients – struggle with the task of retaining given service levels while simultaneously keeping overall costs for bed availability at a minimum. This is because they experience greater variances in patient clientele paired with high emergency admission rates, which significantly reduce planning opportunities. This holds especially true when all beds of a given ward are exclusively assigned to a single department (e.g., Urology, Cardiology, Orthopedics), as there are no predefined alternatives for accommodating overflow patients.

The situation can be improved if several departments share the same beds, i.e., if ward capacities are pooled and patients from different departments are assigned to the same wards. However, not all departments of a hospital can be grouped for medical and social reasons and because of additional costs that would occur if existing wards were combined into larger units. These additional costs, for example, may be caused by further training of nursing staff.

In the present paper, we develop a model and solution approach that solves both the trade-off between savings in bed capacity costs on the one hand and pooling costs that occur on the other hand when combining departments and wards. Aside from this, maximum walking distances between wards and central facilities are minimized for the combinations obtained. In addition,

several constraints are considered when grouping departments and assigning wards to these groups of departments, including medical, social, service, and infrastructural aspects, as well as a varying demand for beds.

We develop a modeling and solution approach for the NP-hard quadratic assignment problem. The approach suggested generates cost-minimal department-ward combinations assuming maximum walking distances for doctors and patients and optimizes the layout of the combinations obtained within the entire hospital.

The modeling and solution approach proposed contributes to existing literature on various distinct aspects. First, the decision problem formulated allows the solution of large, i.e., practical-sized, problem instances, when solving the grouping problem and the layout problem in a lexicographical manner, which is justified by actual management practice. Second, the assignment problem is formulated as a generalized set partitioning problem, which improves the solvability of the problem. The approach therefore allows for time-efficient and exact solutions to this problem. Third, medical, personnel, infrastructural and location constraints are considered including synergy effects regarding additional qualification and management costs incurred as a result of combining departments and wards to pooled capacities. Fourth, we take seasonality effects and differing occupancy distributions per department into account when combining departments. Last but not least, we provide a solution that is highly customizable in terms of department-specific constraints and service levels. Our model has been tested with real data from a large German hospital.

The remainder of the paper is organized as follows. Section 3.2 describes the problem at hand in detail. Section 3.3 reviews the related literature and defines the contribution of the present paper. Section 3.4 then develops the model and the solution approach. Section 3.5 presents numerical results. Finally, Section 3.6 summarizes the key findings and highlights future areas of research.

3.2 Problem Description and Background

The aim of this paper is to develop a model and solution approach designed to group clinical departments of a hospital and exclusively assign existing ward capacity to these groups. Note, that single, i.e., non-grouped departments may also be part of the final solution. A department represents a clinical unit such as Nephrology, or Urology. A ward is a nursing unit, i.e., a defined number of rooms and beds in a specific hallway on a given floor. Several different wards can exist on a single floor. We denote a group of departments (e.g., if Urology and Gastroenterology are put together) as a "department combination". Furthermore, such a combination is termed "department-ward combination" when dedicated ward space is assigned to it (e.g., wards 11 on floor 1 and ward 23 on floor 2 are assigned to the department combination Urology and Gastroenterology). This is visualized in Figure 3.1. Both combinations between departments and between departments and wards are denoted in the literature as clustering, without clearly distinguishing between them.

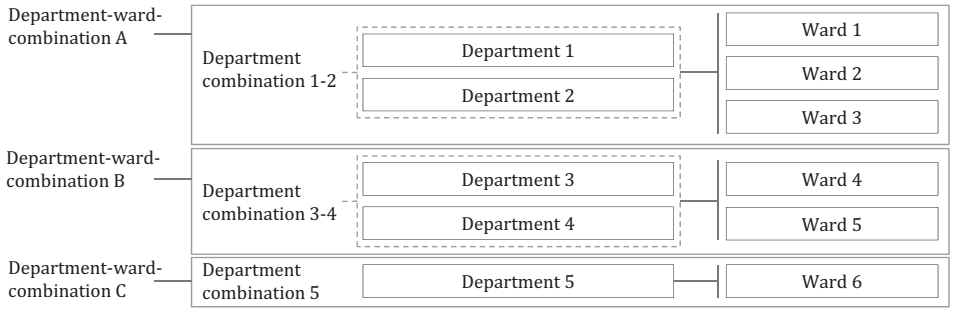


Figure 3.1: Combination of clinical departments and wards

The number of beds required on any given day is driven by a variety of factors. Among others, these include uncontrollable external effects like annual and weekly seasonality or unplanned patient arrivals (especially for emergency patients), and internal parameters like master surgery schedules or personnel rosters and shift plans. The resulting varying demand for beds may either lead to unused capacity or shortages of hospital beds

within a department. Unused capacity results in lower profitability, whereas shortages require significant organizational effort when dealing with overflow situations. This paper therefore investigates the strategic option of improving bed occupancy levels by sharing beds across departments using department-ward combinations instead of exclusively assigning beds of one or several wards to a single department. When departments share wards, this implies that they not only share physical beds, but also the nursing services and resources associated with that bed capacity. Nurses are organizationally allocated to wards (=nursing unit), whereas doctors are typically associated with departments. From an organizational point of view, building department-ward combinations implies that nurses serve patients from different departments within their ward, whereas doctors still only serve patients from their department. However, patients of any one department may be placed in any ward allocated to the respective department combination at hand. This also means that patients from different departments may share one single room.

Two key prerequisites for combining departments are medical and patient compatibility (see Hübner et al. (2016)). Medical compatibility describes the requirement that the risk of dangerous infections has to be kept as low as possible. To that end, certain departments should not be sharing the same ward to avoid mixing their respective patients. For example, immunocompromised patients undergoing cancer treatment should not be exposed to highly infectious patients. Patient compatibility, on the other hand, describes social or emotional requirements of patients. This covers issues such as combining departments with patients with significant differences in severity and type of individual medical condition, age, or gender. For example, it would typically not be permissible to combine a Pediatric department with a Geriatric department.

Focusing on the cost effects when grouping departments, we consider (1) costs for bed availability that can be reduced with larger department com-

binations via pooling effects as well as (2) additional ongoing pooling costs resulting from the increasing complexity and size of the chosen department combinations. In a nutshell, an increasing aggregation level, i.e., combining more departments, decreases the overall bed capacity required and therefore reduces the associated annual costs for providing the respective capacity. On the other hand, the higher the aggregation level, the more complex, i.e., more costly it will be to operate the chosen configuration. We detail these cost effects, which have been jointly developed with physicians, nurses and managers of the case hospital.

(1) Cost Saved via Bed Pooling: Bed occupancy levels of any department are a function of capacity and the number of patient arrivals, their arrival times and lengths of stay. The latter depends on the patient clientele and their respective treatment plans. The number of patient arrivals per day is driven by the stochastic inflow of emergency patients as well as the scheduled appointments for elective patients. Typically, the ratio of emergency and elective patient arrivals varies greatly between departments (e.g., Cardiology has a high share of emergency patients whereas Orthopedics is a heavily elective discipline). Emergency patient arrivals occur randomly and cannot be planned, whereas elective patient arrivals depend on multiple factors, including capacity and availability of operation rooms (OR) as well as medical personnel, and patient preferences. Seasonal effects also play an important role for some departments. For instance, elderly people are more likely to suffer fractures during the winter months when it is slippery and wet outside. Demand during the week may also depend on external factors, such as the daily opening hours of general practitioners. Given these factors, it is clear that different departments will exhibit different occupancy patterns, with varying weekly, monthly and seasonal peaks. Figure 3.2 illustrates a representative case with three departments and two potential department combinations. Combining departments 1 and 3 levels bed requirements better than combining departments 1 and 2.

(2) Additional Ongoing Pooling Costs: However, pooling bed capacity implies that all beds associated with a given department combination have

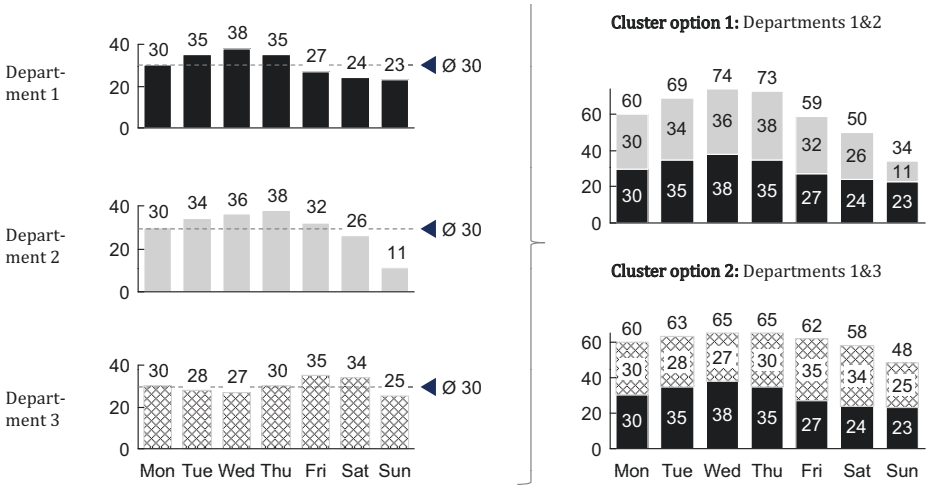


Figure 3.2: Average number of beds occupied during a week (illustrative example)

to be able to accommodate every type of patient from that group. This increases the costs that arise from pooling capacities. There are two cost types:

(2a) Personnel skill-building costs: The nursing staff is the first point of contact for patients. As opposed to doctors who usually see their patients once or twice a day during rounds, nurses have to frequently interact with their patients and are typically assigned to specific wards. Hence, joint ward usage means nursing staff have to cater to the needs of all potential patient types and medical conditions within a specific department combination, and also need to be able to handle emergencies. This requires additional training and qualifications. Skill-building of staff to work in a specific department combination is considered an ongoing effort due to personnel fluctuations and recurring training sessions (e.g., certification processes required to work with specific medical equipment).

(2b) Additional management costs: Larger department combinations (e.g., with more than 200 beds) require more management resources to run day-to-day business efficiently and smoothly than smaller units (e.g., with 20 beds).

This includes operational bed assignment, personnel planning, and so forth. In general, this can only be achieved by providing sufficient resources (e.g., a dedicated bed manager) and appropriate software support (e.g., a bed management system) to create the required transparency across multiple units.

To sum up, pooling effects reduce the number of beds needed for a given service level, which results in lower bed costs. However, the more clinical departments are included in one group, the higher the pooling costs become. Basically, this implies that it is not advisable to combine as many departments into a group as possible just because they are medically compatible. There is a trade-off to be made while seeking to find the lowest-cost solution for this long- and mid-term planning problem.

Alongside minimizing costs, hospital management is also interested in ensuring sufficiently short walking distances for all medical staff serving these department-ward combinations as well as for inpatients. Three stakeholders therefore need to be considered when modeling walking distances, namely nurses, doctors, and patients.

Nurses are generally assigned to a specific ward. Their daily walking distances depend on ward size alone and are not affected by the overall size of the respective department-ward combination.

Doctors are typically assigned to departments, which means that the patients they are attending to may be scattered across all wards within their respective combination. As a result, it seems advisable to avoid heavy scattering of wards within a specific department-ward combination to avoid long walking distances for doctors. However, the amount of time spent while moving between patients is mostly of lower importance given that they typically see their patients once or twice a day while doing rounds. During the rest of their daily routine, doctors work in functional units (e.g., the OR and laboratories) and do office work. Optimizing walking distances

for doctors can therefore be considered a downstream decision as it is mostly driven by the actual operational bed assignment of incoming patients. In practice though, heavy scattering of department-ward combinations should still be avoided wherever possible to ensure coherence for patients and to foster a feeling of togetherness for staff members of a given department. For the long-term planning problem considered it is therefore relevant to ensure that walking distances within any given department-ward combination are not unreasonably high and to consider the objective of minimizing the maximum walking distances between wards as a separately weighted goal.

Inpatients, on the other hand, require services from central facilities, so the distances of all wards assigned to a department combination and the respective central facilities required should be as short as possible. However, equivalent arguments as formulated for doctors suggest this aim should be considered as a separately weighted goal.

To sum up, minimizing walking distances is in general not a primary objective in the given long-term grouping problem. However, it is of particular interest formulating thresholds for maximum walking distances between wards assigned to the same group of departments such that wards of any given group are sufficiently close to each other, and a group is perceived as a single unit by staff and patients. Simultaneously, the maximum distances between any ward of a given department-ward combination and a required central facility should be limited.

This discussion shows the diverse importance of obtaining cost-efficient combinations and short walking distances in hospital practice. Hospital management is usually expected to find the department-ward combinations with the lowest overall operating costs that still adhere to predetermined maximum walking distance thresholds. The minimization of walking distances can thus be considered as a separate goal that is relatively weighted against the occurring operational costs. This leads to a multi-criteria decision problem where the conflicting goals are weighted by the hospital management. This modeling approach also allows solving the entire plan-

ning problem in a lexicographical order, i.e., minimizing costs first, and minimizing walking distances second.

3.3 Related Literature

The key question in bed capacity planning is how to best estimate future occupancy levels. Due to the uncertainty and complexity that hospital situations typically exhibit, there are numerous different approaches determining bed requirements that have been investigated over time. For example, Vassilacopoulos (1985) determines bed requirements based on historical weekly patient data while incorporating constraints such as waiting list length and minimum occupancy levels. Harris (1986) use cyclical OR timetables and length-of-stay data to simulate bed requirements for surgical departments. Alongside simulation techniques, there are numerous papers that utilize queuing theory to plan bed capacity. Green and Ngyuen (2001) for example use queuing models to model patient arrivals and lengths of stay. In a different approach, Cochran and Bharti (2006) and Cochran and Roche (2008) use financial data as well as midnight census data to assess inpatient demand, which they then apply to their simulation to assess bed capacity required.

The result of the bed capacity planning models is mostly that higher occupancy levels are achievable with larger units, i.e., greater pools of jointly utilized bed capacity. Bed pooling approaches are therefore suggested that group clinical departments and assign given wards to these groups. Green and Ngyuen (2001) analyze the bed pooling effects in a large hospital where several surgical specialties have been combined into a single nursing unit. In their report, they emphasize that pooling capacity is especially effective when combining smaller departments, as these typically exhibit high inefficiency due to the underlying stochastic demand. They also acknowledge the importance of taking medical constraints and patient compatibility into consideration when pooling bed capacity. De Bruin et al.

(2010) developed a decision support system based on the Erlang-loss model to determine the appropriate size of wards and expected bed requirements of departments or groups of departments in hospitals. Using historical clinical data of multiple departments, they show that roughly half of the scheduled admissions and almost all emergency admissions can be well described by a Poisson process when considering weekdays and weekends independently. For practical modeling purposes they applied their approach to all departments. However, this implies that all elective patient arrivals are evenly distributed during weekdays, which is not generally the case. As a result, their approach does not differentiate between departments with high or low emergency case ratios, and may therefore underestimate existing weekly seasonality effects. Van Essen et al. (2015) contribute by introducing a methodology to cluster clinical departments and assign wards to these clusters all while not exceeding a pre-specified blocking probability. They propose an approach where they build on the work of De Bruin et al. (2010) and use the Erlang-loss formula to determine bed requirements. They formulate a MIP model that groups departments into department-ward combinations which they call clusters, and allocates ward capacity to these clusters. Their multi-criteria objective function minimizes the maximum number of departments grouped into one cluster and the walking distances between wards assigned to the same cluster, and maximizes the preferences of departments located close to central facilities, e.g., intensive care unit (ICU) and operating room (OR). Since the problem formulated is strongly NP-hard, they propose an approximation model and a two-step-sequential heuristic solution approach. Another related approach has been formulated by Best et al. (2015). They propose an optimization framework for strategic bed capacity layout decisions in large hospitals with the goal of increasing the overall utility for the hospital. Specifically, they try to find a trade-off between forming specialized and non-specialized wings. In their model, non-specialized wings profit from pooling effects while highly specialized wings allow for more focused patient care, which may translate to lower lengths of stay as well as higher utility for patients. Bed requirements are calculated using queuing models without taking seasonal demand into account.

Research Gap and Contribution In general, the suggested bed pooling approaches found in the literature assume Poisson-distributed arrival rates and exponentially distributed lengths of stay. However, this neglects the higher moments of interarrival times and lengths of stay. In particular, departments with a high share of elective patients tend to exhibit less variable interarrival times when comparing real data with Poisson-distributed arrival rates. In addition, the vast majority of approaches neglect demand seasonalities. At most, weekdays and weekends are distinguished. However, occupancy levels of departments demonstrate a distinctive weekly, monthly, and annual seasonality that also varies greatly between departments. Bed pooling approaches should therefore include seasonality effects for relevant cyclic time periods (e.g., week, month, year). We will therefore contribute by both (a) applying empirical distributions of interarrival times as well as lengths of stay, and (b) including department-individual seasonality effects. Furthermore, the current literature on combining clinical departments focuses mostly on maximizing bed occupancy while limiting the aggregation level of departments and satisfying a minimum service level. The literature lacks a comprehensive analysis of actual costs and the trade-off in savings through lower bed costs but higher pooling costs due to greater aggregation of departments. The detailed analysis of decision-relevant costs via a joint project with hospital management constitutes a further contribution of this paper. The suitability of department combinations varies substantially in real life. Simple and generalized approaches to combining departments and assigning ward locations do not meet real life requirements. An approach is required with general and department-specific constraints that can be established individually. This will be integrated in our approach by department-individual constraints. Last but not least, the simultaneous determination of department groups and assignment of wards to these groups is a combination of strongly NP-hard problems (see Van Essen et al. (2015)). An efficient solution approach is therefore necessary to solve scenarios within reasonable time limits. We suggest a new Set Partitioning (SP) modeling approach that can handle large problem settings without having to rely on heuristics as in the other approaches like Van Essen et al. (2015) or Best et al. (2015).

3.4 Model Development

This section develops a model and solution approach to group clinical departments of a hospital and to assign existing wards to these combinations. As denoted in Section 5.2, the underlying problem is to find department-ward combinations that minimize the total costs (i.e., solving the trade-off between cost benefits for pooling bed capacities and the pooling costs incurred) and to minimize distances between wards assigned to one department combination, and between those wards and the central facilities required by the departments of the combination. We propose a decision model that solves the department-ward combination problem by minimizing total costs subject to maximum walking distances and in addition solves the layout problem by minimizing maximum walking distances for the determined department combinations. As minimizing walking distances and minimizing total costs are conflicting goals, we develop a multi-criteria model, which is motivated by managerial practice. The planning problem is formulated as a generalized set partitioning problem (GSPP) designed to find an optimized set of department-ward combinations while considering the trade-off between costs for beds and costs for pooling while respecting distance constraints and a layout that assigns specific wards to department combinations of the entire hospital such that maximum walking distances are minimized.

The remainder of this section is organized as follows. In a first step (3.4.1), we start by describing our general modeling approach – a classical set partitioning problem (SPP) – which we use to group all clinical departments into disjointed subsets where every department is assigned to exactly one subset. In a second step (3.4.2) we then describe the preprocessing required to quantify the data sets and constraints needed to include ward assignment and cost evaluation into the SPP. In a final step (3.4.3) we then formulate the decision model used to solve the entire planning problem defined. The following table summarizes the notation used.

Sets

C	set of feasible department combinations, $C = \{1, 2, ..., c, ..., \|C\|\}$
$C_{\max}$	set of maximal cliques in graph G, $C_{\max} = \{1, 2, ..., i, ..., \|C_{\max}\|\}$
C_{opt}	set of cost-optimal department combinations derived from Model I, $C_{opt} \subseteq C$, $C_{opt} = \{1, 2, ..., c, ..., \|C_{opt}\|\}$
D	set of departments in the hospital, $D = \{1, 2, ..., d, e, ..., \|D\|\}$
$S_{\max,i}$	set of vertices, i.e., departments belonging to maximal clique i of graph G
S_c	subsets of departments belonging to department combination c, $S_c \subseteq D$
$\widehat{S}$	superset of all feasible department combinations, $\widehat{S} = \bigcup_{c \in C} S_c$
W	set of wards in the hospital , $C = \{1, 2, ..., v, w, ..., \|W\|\}$

Parameters

φ_c	cost per department combination c
φ_c^{beds}	annual costs of bed availability for department combination c meeting a predetermined service level
φ_c^{MST}	additional costs per full-time equivalent (FTE) required to operate department combination c
φ_c^{pool}	additional annual pooling costs for department combination c
a_w	number of available beds per ward w
b_c	number of beds required to meet the predefined service level for department combination c
b_d^{FTE}	number of full-time equivalents (FTE) required to operate department d
h_{cw}	distance between ward w and the furthest relevant central facility (e.g., OR) for department combination c
$\tilde{h}_c$	maximum distance allowed between wards and the central facilities of department combination c that are required

Continued on next page

Table 3.1 – *Continued from previous page*

i_{de}	$0-1$ coefficient of incidence matrix I where element $i_{de}=1$ if departments d and e are allowed to be combined into the same department combination; $i_{de}=0$ otherwise
I	$0-1$ incidence matrix I with elements i_{de}, $d,e \in D$
K_{de}	additional personnel skill-building and management costs per FTE (full-time equivalent) when combining departments d and e to a department combination c
l_{dw}	maximum distance between ward w and the central facilities required by department d
Q	$0-1$ coefficient matrix of the set partitioning problem with elements q_{dc}
q_{dc}	$0-1$ coefficient of the set partitioning problem; $q_{dc}=1$ if department d belongs to department combination c; $q_{dc}=0$ otherwise
r_{wv}	distance between wards w and v
$\tilde{r}$	maximum distance permitted between any two wards w,v within any one department-ward combination
R	matrix of the distances between wards with elements r_{wv}, $w,v \in W$
Decision and auxiliary variables	
x_c	$x_c=1$ if department combination c is selected; $x_c=0$ otherwise
y_{cw}	$y_{cw}=1$ if ward w is assigned to department combination c; $y_{cw}=0$ otherwise
z_{wv}	$z_{wv}=1$ if wards w and v are assigned to the same department combination; $z_{wv}=0$ otherwise

Table 3.1: Notation

3.4.1 Overview of General Model Approach

Our modeling and solution approach to build the department-ward combinations is based on a classical set partitioning problem (SPP). The SPP

determines how the items of a given finite set – in our case the set of departments D – can be partitioned into smaller subsets S_c, $c \in C$, i.e., department combinations such that the overall costs are minimized. C represents the set of all department combinations that are medically and socially feasible. The SPP approach requires every department $d \in D$ to be assigned to exactly one and only one partition, while minimizing the overall costs of all chosen department combinations. Modeling the problem as an SPP greatly reduces the combinatorial problem as only $|C|$ potential subsets have to be assessed during runtime as opposed to all theoretical combinations $2^{|D|} - 1$. In essence, the SPP approach is transfering the checking of medical and patient compatibility of department combinations to the preprocessing step (see Subsection (3.4.2) for details). The cost-optimal solution then contains a selection of disjoint subsets $S_c, \forall c \in C_{\text{opt}}$ of the entire set of possible department combinations or subsets. The SPP requires a given set C of $|C|$ subsets S_c, where each item (department) belongs to at least one subset and the costs φ_c associated with each subset S_c are known. In this context the binary decision variable $x_c \in \{0, 1\}$, $\forall\, c \in C$ indicates whether a subset S_c or department combination is part of the solution or not. The SPP can then be modeled as a $0-1$ integer programming model, as formulated below Wolsey and Nemhauser (1999).

Model SPP

$$\min \quad \text{TC} = \sum_{c \in C} \varphi_c \cdot x_c \tag{3.1}$$

subject to

$$\sum_{c \in C} q_{dc} \cdot x_c = 1 \quad \forall\, d \in D \tag{3.2}$$

$$x_c \in \{0, 1\} \quad \forall\, c \in C \tag{3.3}$$

Equation 3.1 minimizes the total costs (TC) of all selected department combinations. Equations 3.2 and 3.3 assure that each department is assigned to one and only one department combination. The parameters $q_{dc} \in \{0, 1\}$, $\forall\, c \in C, d \in D$ define the $0-1$ coefficient matrix Q of the SPP symbolizing whether department d is part of subset S_c , $c \in C$. Based on the SPP we can formulate a generalized SPP that considers additional constraints and the assignment of wards to each of the department combinations chosen.

3.4.2 Preprocessing

The SPP solution approach requires the definition of several sets, subsets and parameters. These sets and figures are defined in the present subsection. Subsection 3.4.2.1 describes how to attain feasible department combinations. Subsequently, the number of beds required (3.4.2.2) and the costs (3.4.2.3) of each of these combinations are quantified. Subsection 3.4.2.4 then defines the distance and location constraints.

3.4.2.1 Feasible Subsets of Departments

For medical and social reasons, not all department combinations are permitted in the problem. Compatibility between departments is defined by the incidence matrix I, where the element $i_{de} = 1$ if departments d and e are allowed to be combined into the same department combination, and $i_{de} = 0$ otherwise. Please note that matrix I is symmetrical, i.e., $i_{de} = i_{ed}$, $\forall\, d, e \in D$ and $i_{d,d} = 1$, $\forall\, d \in D$. All possible and feasible department combinations, i.e., subsets S_c, $c \in C$, can be generated from matrix I. If all departments of a hospital are mutually compatible with each other, then there are $|C| = 2^{|D|} - 1$ possible department combinations. However, in maximum-care hospitals only a limited number of departments are generally allowed to be grouped together. The number of subsets or potentially feasible department combinations, $|C|$, and the size of the $0-1$ coefficient

matrix Q of the generalized SPP will therefore be of quite reasonable size. A branching algorithm is applied using recursive programming to determine the $0-1$ coefficient matrix Q based on the input matrix I. To do this, we represent I as an undirected graph G with $|D|$ vertices, where a direct connection between two vertices d and e stands for compatibility between these two departments. Finding the set of all potential department combinations C is then synonymous with finding all cliques within this graph (see Figure 3.3).

Initialization:	Set $\widehat{S} = \{\}$
Step 1:	Apply Bron-Kerbosh algorithm on graph G
	and determine all maximal cliques $S_{\max,i}$, $i \in C_{\max}$
Step 2:	Create all cliques of graph G
	For all $i \in C_{\max}$ do
	Split $S_{\max,i}$ into all possible cliques with $S_{i,j} \subseteq S_{\max,i}$
	For all $S_{i,j}$; if $S_{i,j} \notin \widehat{S}$; set $\widehat{S} = \widehat{S} \cup S_{i,j}$
Step 3:	Add all single departments to superset $\widehat{S}$
	For all $d \in D$ do; $\widehat{S} = \widehat{S} \cup d$
	Determine set C from superset $\widehat{S} = \bigcup_{c \in C} S_c$
Step 4:	Quantify matrix Q
	For all $c \in C$ and all departments $d \in D$ do
	If $d \in S_c$ set $q_{cd} = 1$; otherwise set $q_{cd} = 0$

Figure 3.3: Algorithm determining the $0-1$ coefficient matrix Q

Step 1 of the algorithm determines all maximal cliques $S_{\max,i}$, $i \in C_{\max}$ within G using the recursive algorithm of Bron and Kerbosch (1973). Step 2 then splits all maximal cliques into all of their respectively possible cliques $S_{i,j}$ and adds those cliques to superset $\widehat{S}$ that are not already part of superset $\widehat{S}$. Step 3 adds all single vertices, i.e., departments to superset $\widehat{S}$. Superset $\widehat{S}$ then contains all feasible department combinations, S_c, $c \in C$, including non-grouped single department setups, $\widehat{S} = \bigcup_{c \in C} S_c$ and set C denotes each feasible department combination, $C = \{1, 2, ..., c, ..., |C|\}$. Finally, the

$0-1$ coefficient matrix Q used in the SPP is generated in step 4 using the superset $\widehat{S}$ obtained in step 3.

3.4.2.2 Bed Requirements per Department Combination

Once the set of potential department combinations is known, the required number of beds to meet a predefined service level has to be determined for each potential department combination. These service levels are either defined by hospital management or set by the legislator. The weekly and seasonal distribution of occupancy levels for every department is assumed to be known. It can reflect historical data as well as expected future demands Hall (2012) Hof et al. (2015). Furthermore, we assume mutually independent distributions of occupancy levels between all departments. Inpatients are typically assigned to one department upon arrival, where they are treated for a specific medical condition, so this assumption is generally satisfied in practice. To calculate the expected occupancy distribution for each department combination $c \in C$ we convolute the probability distribution functions (pdf) of the occupancy levels of the groups or departments independently for individual days within a specified time period (a similar approach is suggested by Fügener et al. (2014)).

Having calculated the pdf for all groups, we can then determine the number of beds b_c needed to meet a predefined service level for every department combination $c \in C$. In our case, it is the day with the highest bed requirements satisfying a predetermined service level within the observed time period (see Figure 3.4). We denote this approach "convolution approach" compared to approaches that only consider the first moments of the respective distributions, such as the Erlang-Loss formula. Please note, that this approach also allows the establishment of individual service level requirements for each department and department combination. In addition, it is important to note that using the day with the highest bed requirement within a predefined time frame (e.g., one week) as the relevant indicator

is a managerial decision. Different approaches are also possible, e.g., the second highest day.

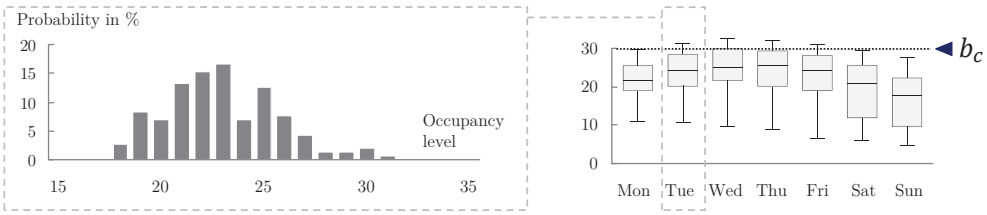


Figure 3.4: Pdf of the daily bed occupancy levels and quantification of the number of required beds, b_c, for department combination c

3.4.2.3 Costs per Department Combination

The overall annual costs φ_c for a potential department combination c, $c \in C$ is the sum of (1) the annual costs for bed availability, φ_c^{beds}, and (2) the annual pooling costs for that combination, φ_c^{pool}.

(1) Costs for beds. The annual costs for beds, φ_c^{beds}, is determined by multiplying the number of beds needed per department combination b_c by the total cost per bed. This comprises not only the cost of the actual bed but more importantly all costs associated with the additional capacity of one bed. Note that these associated costs are to be seen as the ongoing costs needed to have an additional bed, including personnel, infrastructure, utilities, etc., readily available on a given ward. In a non-grouped state, the total number of beds that have to be "readily available" to ensure predefined service levels for each department is higher than in a grouped state that benefits from balancing effects through pooling. Moreover, it is important to note that the required number of beds to be kept readily available is not equivalent to the theoretical total number of beds within all wards associated with a given department or department combination. If rooms in a ward are not required, hospital managers use these for alternative purposes, e.g., waiting room, staff room.

(2) Pooling costs. Combining departments incurs additional pooling costs which also have to be taken into consideration. These additional annual costs per full-time employee (FTE) are predetermined for every combination of any two departments d and e, and symbolized as K_{de}. However, simply adding up the individual costs K_{de} from each pair of departments within a given group will unjustly penalize large groups. Instead we acknowledge synergy effects when it comes to combining departments. These synergy effects can be modeled as a minimum spanning tree (MST) problem where the weights of the edges between nodes, i.e, departments, represent the additional costs per FTE when combining two departments. Using the Prim-Jarnik algorithm Prim (1957) we can quantify the minimum additional costs per FTE required for each department combination c, φ_c^{MST}. Multiplying this cost factor by the total number of required FTE (b_d^{FTE}) of all departments assigned to department combination c, S_c, leads to the annual additional pooling costs:

$$\varphi_c^{\mathrm{pool}} = \sum_{d \in S_c} b_d^{\mathrm{FTE}} \cdot \varphi_c^{\mathrm{MST}} \qquad \forall\ c \in C \tag{3.4}$$

Note that the actual number of patients per department does not change due to the pooling process. The driver for the required FTEs is not the number of beds within one department (or department combination) but the number of patients expected. It is therefore possible to add the number of FTEs b_d^{FTE} theoretically required for every department d in combination c, with $d \in S_c$. In essence, b_d^{FTE} reflects the theoretically perfect workforce size for department d in both the grouped and non-grouped state based on a predefined staff-to-patient ratio.

3.4.2.4 Distances between Wards and to Central Facilities

The set of wards available in the entire hospital is denoted by $w, v \in W$. We assume that the wards are already defined in respect of size, i.e., number of

beds, a_w, and their location in the building. The distance matrix R between each pair of wards, w and v, r_{wv} is calculated from the clinical layout.

In addition, certain departments require certain central facilities, e.g., the ICU or the OR. For doctors and patients the maximum walking distance to these facilities should be limited. Each department d can have multiple requirements of different central facilities. For every potential ward $w \in W$ this relationship is captured prior to preprocessing, where l_{dw} depicts the maximum distance between ward w and the central facilities required by department d. In this context, "required" means that there are sound reasons (e.g., frequent interactions) for the department in question to be close to a specific central facility. As all wards within a department-ward combination can potentially be occupied by patients from any department d within that department combination c, $h_{cw} = \max_{d \in S_c}[l_{dw}]$ denotes the distance between ward w and the furthest relevant central facility for department combination c. Parameter $\tilde{h}_c$ then specifies the maximum distance allowed between wards and the central facilities required for department combination c. The threshold values $\tilde{r}$ and $\tilde{h}_c$ are predefined by hospital management.

3.4.3 Model Formulation

The classical SPP defined in Subsection 3.4.1 can be applied to group departments into department combinations. Along with it the given set of wards of a hospital, W, have to be assigned to the department combinations selected, and additional constraints have to be considered. Model SPP is therefore extended to – what is termed – the generalized set partitioning problem (GSPP) for determining the department-ward combinations.
The resulting multi-criteria model minimizes the total costs TC by choosing the lowest-cost department combinations out of the group of potential combinations and simultaneously minimizes the maximum walking distances $\mathrm{TD}_{\max}$ by assigning wards to these combinations. We introduce α as the

weight for the cost value and β for the distance value since both criteria are measured in different dimensions. The entire model is formulated as follows:

$$\min \qquad \mathrm{Z} = \alpha \cdot \mathrm{TC} + \beta \cdot \mathrm{TD}_{\max} \tag{3.5}$$

subject to

$$\mathrm{TC} = \sum_{c \in C} \varphi_c \cdot x_c \tag{3.6}$$

$$\mathrm{TD}_{\max} = \max_{c,w,v}[\gamma \cdot r_{wv} \cdot z_{wv} + \delta \cdot h_{cw} \cdot y_{cw}] \tag{3.7}$$

$$\sum_{c \in C} q_{dc} \cdot x_c = 1 \quad \forall\ d \in D \tag{3.8}$$

$$y_{cw} \leq x_c \quad \forall\ c \in C;\ w \in W \tag{3.9}$$

$$\sum_{c \in C} y_{cw} \leq 1 \quad \forall\ w \in W \tag{3.10}$$

$$y_{cw} + y_{cv} - z_{wv} \leq 1 \quad \forall\ c \in C;\ w, v \in W,\ w \neq v \tag{3.11}$$

$$h_{cw} \cdot y_{cw} \leq \tilde{h_c} \quad \forall\ c \in C;\ W \in W \tag{3.12}$$

$$r_{wv} \cdot z_{wv} \leq \tilde{r} \quad \forall\ w, v \in W,\ w \neq v \tag{3.13}$$

$$\sum_{w \in W} y_{cw} \cdot a_w - x_c \cdot b_c \geq 0 \quad \forall\ c \in C \tag{3.14}$$

$$x_c \in \{0, 1\} \quad \forall\ c \in C \tag{3.15}$$

$$y_{cw} \in \{0, 1\} \quad \forall\ c \in C;\ w \in W \tag{3.16}$$

$$z_{vw} \in \{0, 1\} \quad \forall\ v, w \in W,\ w \neq v \tag{3.17}$$

As before, the binary decision variable x_c defines whether department combination c is selected or not. The additional binary decision variable y_{cw} defines whether ward w is assigned to department combination c or not. The first part of the objective function (3.5), i.e., equation (3.6), minimizes the total cost TC by choosing the lowest-cost department-ward combinations out of the group of potential department combinations obtained during preprocessing. The second part of the objective function (3.5), i.e., equation (3.7), minimizes the maximum weighted distances $\mathrm{TD}_{\max}$. The first term of equation (3.7) quantifies the walking distance between two wards, w and v, assigned to the same department combination, and the second term specifies the walking distances between the departments to the central facilities they respectively require. Both terms are weighted by the parameters γ and δ, which can be adjusted depending on managerial focus. The decision variable y_{cw} determines the assignment of ward w to group c, and the auxiliary variable z_{wv} defines whether wards w and v are assigned to the same

department combination. Constraint (3.8) ensures that each department $d \in D$ is covered exactly once, where q_{dc} equals 1 if department d is covered by department combination c, and 0 otherwise. Constraint (3.9) denotes that wards $w \in W$ are only assigned to department combinations $c \in C$ that are chosen within the optimal solution. Constraint (3.10) ensures that any ward w can only be connected to one department combination c or not used at all. Thus, it is possible that some wards may be unused and idle afterwards. Constraint (3.11) links the binary decision variables z_{vw} and y_{cw}. Constraint (3.12) limits the distance of any ward w within a department combination c from any central infrastructure required. Constraint (3.13) limits the distance between any two wards w and v assigned to the same department combination to $\tilde{r}$. Constraint (3.14) ensures that the ward capacity assigned covers the number of beds needed, maintaining a predefined service level of department combination c. Finally, constraints (3.15), (3.16), and (3.17) define the binary decision variables.

3.5 Numerical Study

In this section we present detailed results for the approach proposed. We solve a case study for a large maximum-care hospital in Germany in Section 3.5.1. The subsequent Section 3.5.2 investigates our contribution to literature when quantifying bed requirements by considering weekly seasonality and higher moments of interarrival times as well as lengths of stay. Furthermore, we also analyze the general applicability and robustness of the approach suggested. To this end, we present computational experiments using case study data and simulated data that we use to investigate a generalized problem setting as well as a larger hospital scenario in Section 3.5.3. In this final section, we also present and investigate an integrated multi-criteria decision model that simultaneously solves the grouping and layout problem. All preprocessing steps are implemented in Delphi XE4 and Excel 2007. The ILP models are implemented in IBM ILOG Studio v12.6 and solved via CPLEX. All computations were run on a work station

equipped with an Intel Core i7-6700K CPU with 4.0 GHz and 32 GB of RAM.

3.5.1 Case Study

3.5.1.1 Data and Preprocessing

Case Hospital The modeling and solution approach suggested is applied at a large maximum-care hospital in Germany. The case hospital has a total of 22 clinical departments ranging from Pediatric Surgery to Palliative Care. Non-ICU inpatients are accommodated in 36 standard wards, which in the present state are exclusively assigned to one and only one department. The aim of the study was to analyze the feasibility of building department-ward combinations, and to quantify the resulting cost savings when moving from the non-grouped status quo to an optimal grouped state.

Potential Department Combinations In total $|C| = 2^{22} - 1$ distinguishable department combinations would result, when assuming full compatibility between all $|D| = 22$ departments of the case hospital. However, several departments are incompatible for medical and social reasons. We therefore analyzed and defined the bilateral compatibility of each possible combination of two departments during an extensive discussion and several meetings with head doctors, general management, and nursing management. Figure 3.5a) presents the results of this process, i.e., the $0-1$ compatibility matrix I between all departments of the case hospital, where entry 1 indicates the bilateral compatibility of two different departments. Figure 3.5b) shows a resorted version of the original compatibility matrix by applying the approach suggested in Subsection 3.4.2.1 which finds all cliques within the undirected graph defined by the original matrix. The resorted matrix starts with the maximum-sized clique followed by the second largest clique, and so on. Note that the resorted compatibility matrix visualizes some but

not all of the possible and feasible department combinations found by our algorithm. Clique $\{3, 4, 7\}$ for example, which also represents a superset of feasible department combinations, is not immediately obvious from the matrix. A total of $|C|$=115 groups of mutually compatible department combinations were identified by applying the procedure described in Subsection 3.4.2.1. The maximum group size ($\max_{c \in C}[|S_c|]$), i.e., the maximum-sized clique, amounts to six departments.

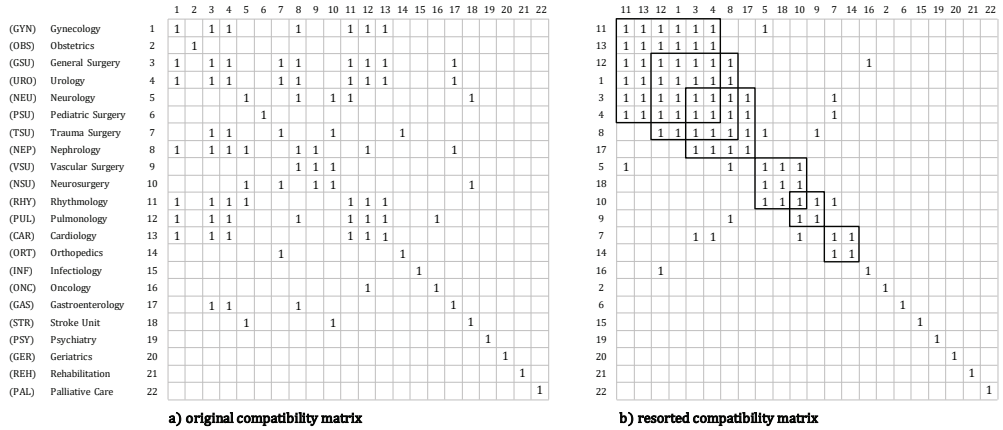


Figure 3.5: Compatibility matrix of departments based on medical constraints and patient compatibility; a) original input from hospital management; b) resorted matrix visualizing a selection of cliques acquired by applying the approach suggested in Subsection 3.4.2.1

Bed Requirements of Department Combinations The expected future occupancy levels per department were created by preparing three years' worth of empirical data based on historical occupancy distributions from 2012 to 2014 and corrected for non-recurring events (e.g., renovation of OR, additional treatments) and trends (e.g., expected case-mix changes). Subsequently, occupancy levels were calculated for each potential department combination for each day of the week independently. Bed capacity requirements b_c were then derived from this data for each potential department combination by requiring a given unified service level of 95% based on the weekday with the highest anticipated occupancy level. Note that this can be a different day for each department combination depending on the composition of one group.

Costs of Department Combinations For every combination of two departments, nursing management has defined a cost factor per FTE, K_{de}, that quantifies the additional ongoing pooling costs due to higher management complexity and continuous skill-building requirements if these two departments were to be combined. Using this data set, we then calculated the respective additional pooling costs φ_c^{pool} for each potential department combination $c \in C$. Finally, the total costs per department combination were determined by $\varphi_c = \varphi_c^{\text{beds}} + \varphi_c^{\text{pool}}$.

Distances Between Wards In the case hospital the available wards are distributed across five floors. Each floor consists of four wings, with each wing consisting of two wards. This results in 40 wards, with four wards that are used for central facilities. Wards differ in size due to architectural restrictions or partial usage for other functions. In the case study we therefore consider 36 wards with a capacity ranging from 8 to 24 beds per ward. Furthermore, we only consider standard wards as potential capacity for inpatients of combined departments. Specialized wards such as the intensive care unit (ICU) are modeled as central facilities that do not share beds with standard wards. For simplification purposes we used a normalized distance unit representing the distance between two wards w and v, r_{wv}. Moving from floor to floor amounts to three units. Moving between wings on a single floor takes one to three units depending on the respective distances between these wings. Moving between wards of the same wing amounts to zero units. Figure 3.6 illustrates the respective distances on a single floor at the case hospital with four wings (a, b, c, and d) and eight wards.

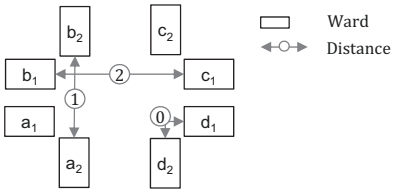


Figure 3.6: Relative distances on a single floor between wards in the case example

For example, the distance between wings a and b is 1 unit, between b and c is 2 units, and between a and c is 1+2=3 units. Specific location constraints

per department were jointly defined with nursing management and head doctors. Accordingly, the maximum distance within a department-ward combination was defined as $\tilde{r} = 10$ distance units.

Distances to Relevant Central Facilities Depending on the department, there may be either zero, one, or multiple location requirements regarding central facilities. For example, all wards of the Obstetrics department have to be close to the laboratory and delivery rooms as well as to the OR to perform C-sections, whereas the wards of the Oncology department can be located anywhere within the hospital. Given this data per department, we can determine the distance h_{cw} to the furthest central facility for every possible assignment of a department combination c with any single ward w. The maximum distance to a relevant central facility is limited by hospital management to $\tilde{h}_c = 10$ distance units.

Relevance of different criteria Hospital management prioritizes generating cost benefits from bed pooling before minimizing walking distances for patients and doctors. Finding the cost-minimal set of department combinations is defined as the primary goal as long as walking distances are within defined boundaries. Minimizing walking distances for patients and doctors is then formulated as a subordinate goal. Model GSPP is therefore solved in a lexicographical manner, i.e., where $\alpha/\beta \geq 10.000$, such that the focus of the model lies in generating a cost-minimal set of department combinations for which walking distances are minimized simultaneously. Note, however, that this approach does not bypass any preset maximum walking distance thresholds for the chosen set of department combinations. In addition, the hospital management considers occurring walking distances between wards assigned to the same department combinations much more unpleasant than possible distances to any central facilities required. This implies that equation (3.7) is therefore implemented using $\gamma \gg \delta$.

3.5.1.2 Results for the Case Hospital

Given all cost, bed, and distance data for each potential department combination, we apply the model. This results in $C_{\text{opt}} = 16$ disjoint groups of department combinations. One of these groups contains three departments, four groups contain two departments, and 11 departments remain isolated (see Figure 3.7). Compared to the present organization of the hospital,

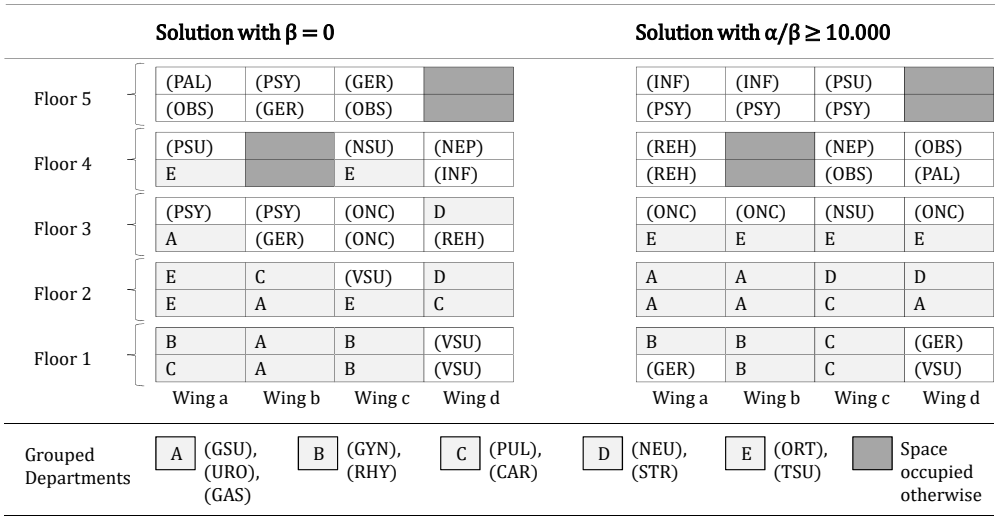


Figure 3.7: Comparison of resulting layouts after combining departments and wards using different settings for α and β

50% of all departments were grouped into department combinations with an average size of 2.2 departments within one group. As a result of this grouping, the number of beds needed to maintain a unified service level of 95% for all departments was reduced by 3.3%, while total costs were cut by 2.1%. The resulting ward assignment can also be seen in Figure 3.7. To verify that the application of high ratios of α/β actually leads to a cost-optimal grouping of departments we first ran our model without walking distance optimization, i.e., with $\beta = 0$. When comparing this result with the solution of the model using $\alpha/\beta \geq 10.000$, it can be seen that even the cost-optimal set of department-combinations allows for very good results regarding walking distance optimization. In the case considered, the maximum distances between wards assigned to the same group of de-

partments were reduced to three distance units while keeping the most cost-efficient set of department combinations.

3.5.1.3 Sensitivity Analyses of Cost Parameters

Management typically requests sensitivity analyses of the model parameters applied since some parameters may be difficult to quantify or are not as fixed as formulated by the right side of a "hard" constraint. The cost parameter φ_c for example is highly dependent on the relationship between the annual costs for beds, φ_c^{beds}, and the annual pooling costs, φ_c^{pool}, of department combination c. In practice, setting these two cost factors may become challenging as there are many underlying aspects that either cannot be determined easily or are highly dependent on subjective valuation. For example, the exact cost of holding one bed or room available including personnel, infrastructure, etc. is difficult to pinpoint. It is therefore beneficial to evaluate the solutions of the optimization model along different relationships of both cost parameters. In this setting we define the annual pooling costs as percentage f, $f \in [0\%,\ 100\%]$ of the additional annual cost savings via bed pooling $\varphi_c^{\text{pool}} = f \cdot \left[\sum_{d \in S_c} \varphi_d^{\text{beds}} - \varphi_c^{\text{beds}}\right] = f \cdot \Delta\ \varphi_c^{\text{beds}}, \forall\ c \in C$. Here, $\Delta\ \varphi_c^{\text{beds}}$ depicts the cost-savings for combination c that occur as a result of pooling ward capacity. Specifically, φ_d^{beds} represents the number of beds required for an individual department d in the non-grouped state. Figure 3.8 then depicts the grouping results achieved when factor f varies between 0% and 100%.

Setting $f = 0\%$ means that no additional pooling costs emerge when grouping departments, and clinical departments will be grouped as much as possible. In this case 13 out of 22 departments are grouped into one of the department combinations generated. However, setting $f = 100\%$ leads to prohibitively high grouping costs, with the result that none of the 22 departments are combined at all. Note that the solution generated for the example case, i.e., $C_{\text{opt}} = 11$, is remarkably stable over a wide range of cost

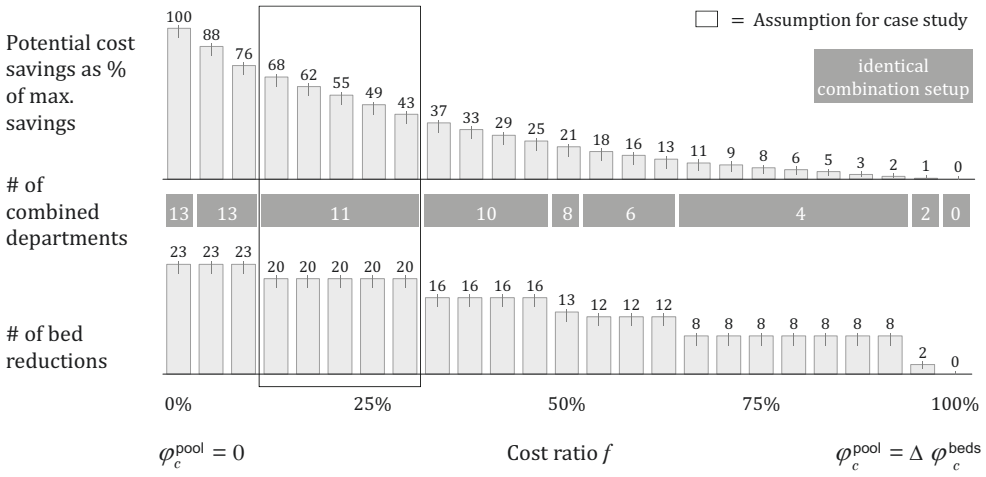


Figure 3.8: Sensitivity analysis of cost ratio $f = \frac{\varphi_c^{\text{pool}}}{\Delta\ \varphi_c^{\text{beds}}}$

ratios, i.e., $f \in [16\%,\ 32\%]$. In the present case, a relatively robust solution is therefore found with respect to the grouped departments.

3.5.2 Quantifying Bed Requirements

Current literature on department grouping does not leverage weekly seasonality, and uses average data of a week to estimate bed requirements. In addition, higher moments of interarrival times and lengths of stay are generally neglected. We would like to contribute to the literature by proposing a different approach to quantify bed requirements. We will show in Section 3.5.2.1 that seasonality matters, and in Section 3.5.2.2 that empirical distributions are better suited for estimating bed occupancy levels than the Erlang-Loss formula.

3.5.2.1 Impact of Weekly Seasonality

In general there is no doubt that occupancy levels of clinical departments exhibit a strong weekly seasonality that becomes especially visible when weekdays and weekends are compared. Because of this observation several approaches suggested in the literature either average the occupancy levels on weekdays or use the peak occupancy level during the week when quantifying the bed capacity required per department (see De Bruin et al. (2010) or Van Essen et al. (2015)). However, the detailed occupancy pattern on each day of the entire week has not yet been considered, which becomes especially relevant when quantifying bed requirements of grouped departments (see Section 5.2). To obtain more insights on how incorporating seasonality effects impacts bed requirements and potential overflow, we investigate three different approaches, all of which use the convolution approach to calculate bed requirements:

- Approach 1: Omitting weekly seasonality entirely and quantifying an aggregated occupancy distribution for each department on a weekly basis.
- Approach 2: Excluding weekend data and quantifying an aggregated occupancy distribution for each department on a weekday basis.
- Approach 3: Considering weekly seasonality and explicitly modeling the occupancy distribution for each day of a week, i.e., the approach proposed in this paper.

We again use the historical occupancy levels of three years obtained from our case hospital. Table 3.2 summarizes the resulting department combinations and patient overflows. Note that we list only the grouped departments.

The results convey three distinct insights. First, changing the approach for calculating bed requirements also impacts the resulting department combinations. For example, the group comprising both PUL and CAR without any other department has only been chosen in approach 3. In other words, not including seasonality effects penalizes complimentary occupancy patterns. Second, assuming average occupancy rates across

Approach 1: No seasonality		Approach 2: Partial seasonality		Approach 3: Full seasonality	
Combinations	Overflow in %[1]	Combinations	Overflow in %	Combinations	Overflow in %
GSU,GAS	3.5	GSU,GAS	1.7	-	-
-	-	-	-	GSU,GAS,URO	1.6
-	-	GYN,RHY	2.1	GYN,RHY	2.1
GYN,URO	9.6	-	-	-	-
NEP,VSU	8.5	NEP,VSU	3.6	-	-
NEU,STR	4.4	NEU,STR	4.4	NEU,STR	1.2
-	-	-	-	PUL,CAR	1.7
RHY,PUL,CAR	13.1	-	-	-	-
TSU,ORT	5.4	TSU,ORT	3.5	TSU,ORT	2.1
-	-	URO,PUL,CAR	6.2	-	-

[1] Amount of days with capacity overflow

Table 3.2: Analyses of seasonality effects when grouping departments and assigning wards

the week or during weekdays may lead to unwanted overflow effects, as can be seen in the results of approaches 1 and 2. Third, super-positioning multiple departments that exhibit similar peak days during the week strongly affects the overall occupancy level variance of the resulting combination. This becomes especially visible when applying approach 1, where overflow probabilities per combination range from 3.5% to 13.1%.

3.5.2.2 Impact of Distributional Assumptions

When modeling the distribution of occupancy levels of clinical departments, the question arises: should one apply empirical or theoretical distributions? Most approaches suggested in the literature rely on theoretical distributions since these approaches mainly apply queuing theory models, e.g., the Erlang-loss formula, when quantifying bed requirements (see De Bruin et al. (2010) Best et al. (2015) Van Essen et al. (2015)). However, in our modeling and solution approach we propose using empirical occupancy data corrected for trends and non-recurring events. Achieving some insights into the above mentioned question, we also compute the bed capacity required using the Erlang-loss formula. Table 3.3 shows the results achieved for the grouped departments. The "convolution approach" equals the proposed approach in this paper by convoluting discrete empirical distributions when quantifying the bed capacity required for each department combination, b_c, $\forall c \in C$. The

second approach, however, determines bed requirements by the Erlang-loss formula using only weekday occupancy levels.

Convolution Approach:			**Erlang-Loss Approach:**		
Combinations	**Overflow** in %[1]	b_c	**Combinations**	**Overflow** in %[1]	b_c
GSU,URO,GAS	1.6	91	GSU,URO,GAS	6.4	88
GYN,RHY	2.1	52	-	-	-
-	-	-	GYN,RHY,CAR	12.9	83
-	-	-	NEP,VSU	0.1	37
-	-	-	NEU,NSU,STR	0.0	66
NEU,STR	1.2	43	-	-	-
PUL,CAR	1.7	56	-	-	-
-	-	-	PUL,ONC	1.4	52
TSU,ORT	2.1	85	TSU,ORT	12.8	77
Comparison of bed requirements to meet target service level					
Uncombined state	$\sum b_c = 611$		Uncombined state	$\sum b_c = 643$	
Combined state	$\sum b_c = 591$		Combined state	$\sum b_c = 592$	

[1] Amount of days with capacity overflow

Table 3.3: Empirical vs. theoretical occupancy distributions

The comparison between these two approaches shows that convoluting empirical occupancy distributions allows for a more precise approximation of the beds required to fulfill the predefined service level of combined departments. Note that using different approaches to calculating b_c inevitably leads to the model choosing partly different department combinations. In essence, our analyses show that the assumption of the Erlang-loss formula, i.e., a Poisson-distributed arrival process and exponentially distributed lengths of stay, overestimates bed requirements for smaller department combinations or single departments, particularly if these departments mostly treat elective patients and therefore exhibit a relatively low rate of randomly distributed emergency patient arrivals.

3.5.3 Trade-off between Cost and Walking Distance Optimization

3.5.3.1 Results for the Original Case Study Data

In the following we compare the results achieved by Model GSPP when modifying the ratio α/β, assuming the original data of the case study. Table 3.4 presents the results.

Setting α/β	Run time;[1] *in sec.*	Setup[2]	Combinations (depart. grouped) *total number*	Beds pooled, *in %*	Cost savings;[3] *in %*	Maximum ward-ward distance	Maximum ward-infra. distance
10,000	5.8	a	5 (11)	54	100	3	21
1,000	4.5	a	5 (11)	54	100	3	21
100	5.0	a	5 (11)	54	100	3	21
10	17.3	b	5 (10)	51	94	1	22
1	46.9	b	5 (10)	51	94	1	22
0.1	62.0	b	5 (10)	51	94	1	22
0.01	152.3	c	4 (9)	40	67	1	19
0.001	249.7	c	4 (9)	40	67	1	19
0.0001	130.7	c	4 (9)	40	67	1	19

[1] Computational time of integrated model
[2] Each letter describes a unique department-ward-layout combination;
[3] Normalized cost savings with regard to non-clustered status quo; 100% = maximum costs savings achieved by hierarchical approach, i.e., $alpha/beta \geq 10,000$

Table 3.4: Analysis of integrated approach for varying weight ratios, based on case study data

The results of Table 3.4 reveal that the case example can be solved within a reasonable computation time for all α/β-ratios. The required computation time, however, gets longer when the ratio α/β decreases. Second, an increase in the importance of walking distances within the integrated model, i.e., a declining α/β-ratio, leads to new sets of department combinations. This effect shows the rising impact of layout decisions on the chosen set of department combinations when moving out of a setting with a dominant focus on costs. Thus, assigning higher weights to distance optimization takes a toll on pooling benefits. Third, the level of grouping, i.e., the number of departments and beds in a grouped state, declines with a rising importance of walking distances in the integrated model. This is in line with what is to be expected. However, the real insight here is the fact that the improvement in maximum ward to infrastructure distances is rather small compared to a significant loss of pooled beds when increasing the walking distance weight β. Furthermore, it is important to note, that even when

assuming very high α/β-ratios, i.e., $\alpha/\beta \geq 10,000$, resulting in the "setup a", the maximum distance between two wards of the same combination only amounts to 3 normalized distance units (see Figure 3.6), which for a typical hospital setting would be more than acceptable given the overall size of the hospital.

3.5.3.2 Variation of Layout Restrictions

To build a generalized case, we have replaced the original department-specific ward-to-ward and ward-to-infrastructure threshold restrictions with unified distance thresholds for every department. We then investigated five additional and different settings for thresholds ranging from strong constraints (termed as T1) to weak constraints (T4) and no distance constraints (no T). As thresholds we use 6, 8, 10, and 12 units for the ward-ward distances and 18, 20, 22, and 25 for the ward-infrastructure distances, respectively. Table 3.5 summarizes the results.

	T1		T2		T3		T4		no T	
α/β	Time	Setup	Time	Setup	Time	Setup	Time	Setup	Time	Setup
10,000	9.5	e	7.7	a	14.1	a	17.5	a	20.8	a
1,000	7.2	e	8.2	a	12.1	a	20.1	a	20.3	a
100	10.4	e	7.9	a	10.6	a	18.0	a	15.2	a
10	7.0	e	8.6	h	11.8	a	21.8	a	26.2	a
1	16.9	f	10.4	h	32.5	a	46.3	a	99.5	a
0.1	36.3	f	47.0	i	56.2	a	175.9	a	641.8	a
0.01	15.1	f	65.7	c	222.2	c	550.3	c	986.9	c
0.001	17.4	f	135.4	c	115.6	c	334.8	c	753.8	c
0.0001	21.9	f	90.1	c	247.6	c	427.0	c	600.3	c

Runtime of CPLEX solver, *in seconds*; each letter in columns "setup" describes a unique department-ward-layout combination

Table 3.5: Analysis of computational efficiency with different walking distance thresholds

The key conclusions drawn for our case study (see Table 3.4) hold true with the extended analysis on the distance thresholds (see Table 3.5). Additionally, it becomes clear, as one would assume, that predetermined walking distance thresholds impact computation time, with looser boundaries leading to a wider solution space and thus a higher runtime. Moreover, there is also a constraining effect visible with regard to the department combination setup when applying very strong distance boundaries. For instance, even when assuming $\alpha \gg \beta$ the model does not allow to select the cost-minimal

"setup a" when applying very strict distance thresholds. Finally, the results in Table 3.5 also show that even when removing the initial walking distance thresholds entirely, the computation time remains within a reasonable time frame for the case hospital setting.

3.5.3.3 Larger Data Sets

The modeling approach suggested minimizes the maximum walking distances (see equation (3.7)). This criteria has been derived from our work with the management of the case hospital. In a further generalization, we introduce an alternative objective function that minimizes the sum of all walking distances. In particular, we compare the effect of these two criteria for evaluating the walking distances. For this purpose the objective function (3.5) of the integrated model is reformulated as follows:

$$\text{min! } Z = \alpha \cdot \sum_{c \in C} \varphi_c \cdot x_c + \beta \cdot (\gamma \cdot \sum_{w \in W} \sum_{v \in W} r_{wv} \cdot z_{wv} + \delta \cdot \sum_{c \in C} \sum_{w \in W} h_{cw} \cdot y_{cw}) \quad (3.18)$$

We further analyze the scalability of our approach beyond the case study setting for the maximum-care hospital, which was already large. In so doing, we orientated ourselves to the largest hospitals available in Europe when constructing the additional examples. These large hospitals have a maximum of 40 departments in a single building complex. Additionally, we assume that typically no more than 10 departments can be combined into the same group due to compatibility constraints. To this end, we have constructed a fictitious hospital building with 65 wards spread across 7 floors comprising a total of 1,396 standard beds, i.e., excluding intensive care units, intermediate care units, and non-stationary emergency units. Such a capacity of standard inpatient beds in one single building is comparable to the largest European hospitals. This is termed "very large hospital" (VLH). We then randomly determined all necessary parameters such as

bed requirements, costs and so forth while applying reasonable boundaries to keep the scenario realistic. Specifically, we built fictitious departments and department combinations and based their respective bed requirements, distance thresholds and cost parameters on experience from our case study by applying similar distributions, minimums, and maximums. This leads to 960 potential department combinations comprising at least two departments each.

We analyze the VLH case using different walking distance thresholds. The thresholds applied are equivalent to the ones used in the generalized case study. In order to showcase only the extreme settings in this simulated case, we limit ourselves to the following three scenarios, including strong thresholds (T1), weak thresholds (T4), as well as no thresholds (no T). Furthermore, we compare these threshold-scenarios using the two distance objectives (*MinMax* and *MinSum*).

	'MinMax'						*'MinSum'*					
	T1		*T4*		*no T*		*T1*		*T4*		*no T*	
α/β	Time	Setup	Time	Setup	Time	Setup	Time	Setup	Time	Setup	Time	Setup
10,000	294	A	1,217	B	2,779	B	317	A	1,519	B	3,783	B
1,000	325	A	844	B	1,356	B	5,227	A	>24h	N/A	>24h	N/A
100	327	A	924	B	2,071	B	>24h	N/A	>24h	N/A	>24h	N/A
10	721	A	30,594	C	>24h	N/A	>24h	N/A	>24h	N/A	>24h	N/A
1	5,633	A	>24h	N/A	>24h	N/A	>24h	N/A	>24h	N/A	>24h	N/A
0.1	>24h	N/A	>24h	N/A	>24h	N/A	>24h	N/A	>24h	N/A	>24h	N/A
0.01	>24h	N/A	>24h	N/A	>24h	N/A	>24h	N/A	>24h	N/A	>24h	N/A
0.001	>24h	N/A	>24h	N/A	>24h	N/A	>24h	N/A	>24h	N/A	>24h	N/A
0.0001	>24h	N/A	>24h	N/A	>24h	N/A	>24h	N/A	>24h	N/A	>24h	N/A

Runtime of CPLEX solver, *in seconds*
Each letter in columns "setup" describes a unique department-ward-layout combination

Table 3.6: Analysis of the computational efficiency of the very large hospital case

The additional set of analyses for the VLH-case delivers three insights. First, using large α/β-ratios , i.e., $\alpha/\beta \geq 10,000$ has a very positive effect on computational times, as expected. Second, the proposed *MinMax* objective performs better throughout the different parameter settings. Due to the large size of the simulated problem instance, the solution time for the VLH-case increases exponentially when the α/β-ratio decreases within the integrated model. This may lead to problem instances that cannot be solved in acceptable time limits, especially when using the *MinSum* objective. Last but not least, the effect of walking distance thresholds proves to be a lot stronger in the VLH-case compared to the standard-sized generalized case

study, which is also to be expected given the large size of the simulated VLH case.

All in all, the results of our experiments show a general applicability for our model for a generalized hospital case with varying parameters. Moreover, a "hierarchical setting", which is motivated by managerial practice allows optimal solutions to be found for the department-ward combination problem even for very large hospital settings within very short computational time frames. The modeling and solution approach suggested therefore seems to be well suited as a decision-support model in real-life hospital settings since the approach allows in-depth sensitivity analyses of different setups and parameters.

3.6 Conclusion and Further Areas of Research

The present paper has considered the problem of grouping clinical departments of a maximum-care hospital and assigning the available wards of the hospital to those groups. The entire problem is denoted as a strategic department-ward combination problem. We developed a novel modeling and solution approach based on a generalized set partitioning formulation. Using a real-life case study, we demonstrated that the proposed approach is generally applicable and leads to significant cost savings while ensuring a minimum acceptance level of inpatients. Decision-relevant costs for bed availability and pooling costs as well as maximum walking distances are considered when creating department-ward combinations. In alignment with the literature we found that the weekly demand seasonality has to be considered when quantifying bed requirements of department combinations. It is however insufficient to only distinguish between weekdays and weekends. Department-specific daily demand patterns have to be taken into consideration. In addition it is shown that the Erlang-loss formula

leads to imprecise approximations for bed requirements. Using empirical distributions is therefore much more advisable. For the case hospital considered, we demonstrated that moving from an ungrouped status quo to an appropriate grouped state with pooled ward capacity leads to significant cost savings, lower bed requirements and higher bed availability.

Nevertheless, our proposed modeling and solution approach offers several areas for future research. Due to the long-term planning horizon of the problem at hand we neglect the transition process as well as one-time costs as we are only analyzing long-term steady states. Further research should focus on how to manage the transition process as well as finding a cost-efficient implementation sequence. Seasonally driven demand variation could also be an interesting area of future research. In the present paper, we focused on weekly seasonality effects, but we disregarded annual seasonality effects. Here, one could imagine seasonal closures or temporal reallocations of selected ward capacity. It is also crucial to understand and evaluate the impact of tactical and operational decisions within a hospital on the generally strategic department-ward combination problem. For example, there are mutual dependencies between master surgery scheduling, personnel rostering and bed occupancy planning. Typically, these interdependencies are modeled by including respective assumptions and constraints and solving one problem at a time. Future research should focus on shedding more light on the relationships of different hierarchical planning problems. In particular, it would be very beneficial to understand the impact of changing the sequence of these planning problems on overall bed occupancy. Here, one of the key questions could be how to integrate long-term OR scheduling. Many approaches (see Fügener et al. (2014)) exist that investigate the effect of OR scheduling on downstream units, but the strategic effect on the department-ward combination problem has not yet been thoroughly examined to our knowledge. We focus on the department-ward combination problem and solve the cost-minimization and distance-minimization problem. However, we assume a given pool of wards which are pre-specified in terms of location and size as well as central facilities that are already firmly located. These assumptions may be relaxed in the future, meaning that the

entire problem then becomes harder to solve (see e.g., Helber et al. (2015)). Finally, it would be fruitful to undertake research on how to handle tactical and operational bed allocations within department combinations, and how to handle overflow situations. Having an effective decision support system to allocate beds to patients on a daily basis would be necessary in order to effectively manage large department combinations and exploit synergies from pooled bed capacity.

4 Operational Patient-Bed Assignment Problem in Large Hospital Settings including Overflow and Uncertainty Management

Co-authors: Fabian Schäfer, Alexander Hübner, Heinrich Kuhn
Published in *Flexible Services and Manufacturing Journal* on 10 January 2019

Abstract Managing patient to bed allocations is an everyday task in hospitals which in recent years has moved into focus due to a general rise in occupancy levels and the resulting need to efficiently manage tight hospital bed-capacities. This holds true especially when being faced with high volatility and uncertainty regarding patient arrivals and lengths of stay. In our work with a large German hospital we identified three main stakeholders, namely patients, nurses, and doctors, whose individual objectives and constraints regarding patient-bed allocation (PBA) lead to a potential trade-off situation. We developed a decision support model that tackles the PBA problem considering this trade-off, while also being capable of handling overflow situations. In addition, we anticipate emergency patient arrivals based on historical probability distributions and account for uncertainty regarding patient arrival and discharge dates. We develop a greedy look-ahead heuristic which allows for generating solutions for large real-life operational planning situations involving high ratios of emergency patients. We demonstrate the performance of our heuristic approach by comparison with the results of a near-optimal solution achieved by Gurobi's MIP solver. Finally, we tested our approach using data sets from the literature as well as actual clinic data from our case study hospital, for which we were able to reduce overflow by over 96% while increasing overall utilization by 5%.

4.1 Introduction

This paper deals with the operational planning question of assigning incoming patients to specific rooms and beds upon their arrival at the hospital. This so-called patient-bed allocation (PBA) problem has been gaining more and more attention in recent years after a basic version of the problem was formulated by Demeester et al. (2010). Based on this seminal work, related research was mostly directed at either improving the computational efficiency (see for example Bilgin et al. (2012) or Range et al. (2014)) or proposed ways to incorporate upstream planning problems such as surgery or elective patient scheduling (see for example Ceschia and Schaerf (2016)).

In our joint project with a large German hospital we identified several challenges with respect to the PBA problem that have to be dealt with in real-life situations including emergency and elective patients of all major disciplines.

First and foremost, large hospitals with 500 or more beds covering all major disciplines exhibit high ratios of emergency patients, e.g., up to 90% in internal disciplines such as cardiology and gastroenterology. Due to the nature of the patient clientele in these hospitals (inherent multimorbidity, unknown medical history, etc.) it is oftentimes not possible to accurately determine the actual length of stay (LOS) of a patient once they arrive as well as throughout their stay.

Second, a shift in demographics as well as advances in medical technologies are forcing hospitals to operate as cost-efficiently as possible. This leads to high overall bed occupancy levels which in turn may more often lead to situations in which bed capacities are insufficient. To minimize such overflow situations while keeping bed occupancy levels high, a common approach is to pool bed capacities across similar medical disciplines to create a balancing effect across the associated wards (see for example Hübner et al. (2016) and (2018)). However, as opposed to single wards with ten to

twenty beds, managing operational patient bed assignments within a set of designated wards comprising more than a hundred beds leads to a highly complex planning problem which typically cannot be dealt with efficiently by conventional planning approaches, e.g., a dedicated bed planner who manually assigns patients to beds.

Third, there is a need to adapt patient-bed allocations ad-hoc to changes, as any plan made at a certain point in time is likely to be obsolete only a few hours later due to new emergency arrivals, sudden complications after surgery, or new diagnostic findings (see for example Hulshof et al. (2016)). In practice, this means that the decision problem has to be solved whenever there is a change in the system which merits the physical allocation of a newly arrived patient or a patient waiting in an overflow area to a bed. The large hospitals considered in this paper are deciding on this issue several hundred times a day.

Fourth, three major stakeholders have to be kept in mind, namely patients, nurses, and doctors. Specifically, it is important to make the stay for patients as comfortable as possible while simultaneously respecting patient-specific constraints, balancing the workload for nurses, and making it as efficient as possible for doctors to do rounds.

In essence, this leads to an assignment problem that respects the diverse interests of patients, nurses, doctors and hospital management while simultaneously considering medical, gender, and capacity constraints. Hence, there is a need for a PBA system which is capable of anticipating future developments while at the same time being able to provide quick online recommendations for patient-bed allocations within seconds when prompted (see for example Hulshof et al. (2012)).

In this regard, the present paper proposes a new modeling and solution approach to the PBA problem that incorporates stakeholder-specific objectives for patients, nurses, and doctors. In addition, the paper provides a greedy look-ahead heuristic that allows for flexible bed allocations while

managing overflow situations and anticipating future arrivals of elective and emergency patients. The findings and insights discussed herein are not limited to the German health-care system but may well be of importance to any large hospital setting faced with the above-described circumstances.

The remainder of this paper is structured as follows: Section 4.2 provides a detailed problem description discussing relevant literature and further elaborates on the specific contribution of this paper. Section 4.3 lays out the modeling and solution approach. Section 4.4 then provides numerical examples. In particular, we compare the results of our heuristic solution approach with the results of a near-optimal solution achieved by Gurobi's MIP solver for selected problem instances. Furthermore, we test our approach with real-life data from a large hospital in Germany and use several sensitivity analyses to investigate solution quality and run time required. In addition, we further test our approach with data from the literature (see Demeester et al. (2010)). Finally, Section 4.5 presents a summary of the main results and gives an outlook on possible future avenues of research.

4.2 Problem description, related literature and contribution

To understand the main objectives of the patient-bed allocation process in a hospital we interviewed nurses, doctors, and hospital management of our case-hospital. The following subsections describe the general planning problem and related literature as well as open research questions that we tackle.

4.2.1 General planning problem

Operational bed occupancy management in hospitals comprises two inherently different planning problems, namely patient admission scheduling (PAS) and patient-bed allocation (PBA). It should be noted, that in the literature these expressions have been used with varying definitions. We consider the PAS problem as merely comprising the problem of scheduling elective patient admission dates. The PBA problem, however, relates to the problem of allocating a physical room and bed to a patient. In large hospitals with more than 500 beds and a high rate of emergency arrivals the two decision problems are typically solved in a hierarchical manner for reasons set out below.

In a first step, the goal of a PAS system is to ensure a high and balanced utilization of the available bed capacity over time. In principle, four patient classes need to be considered. Namely, elective patients and emergency patients who are already physically available in the hospital, as well as planned elective patients and future emergency patients who are already scheduled to or anticipated to arrive in the future, respectively. Figure 4.1 shows a schematic example for a typical PAS situation and depicts the number of beds occupied by or reserved for the afore-mentioned patient classes for the first night of the planning horizon, i.e., a Monday night, and on each of the consecutive 13 nights.

On the first Monday a certain number of beds are already physically occupied by elective and emergency patients. These numbers decrease over time as most patients occupying a bed on the first day of the planning horizon will leave the hospital on the following days. Note, these numbers mostly stay stable on Saturdays and Sundays, since no discharges take place on those days.

In addition, a certain number of beds have to be reserved for incoming elective and emergency inpatients which are planned or anticipated to show up in the future. Whatever bed capacity is still available after incorporating

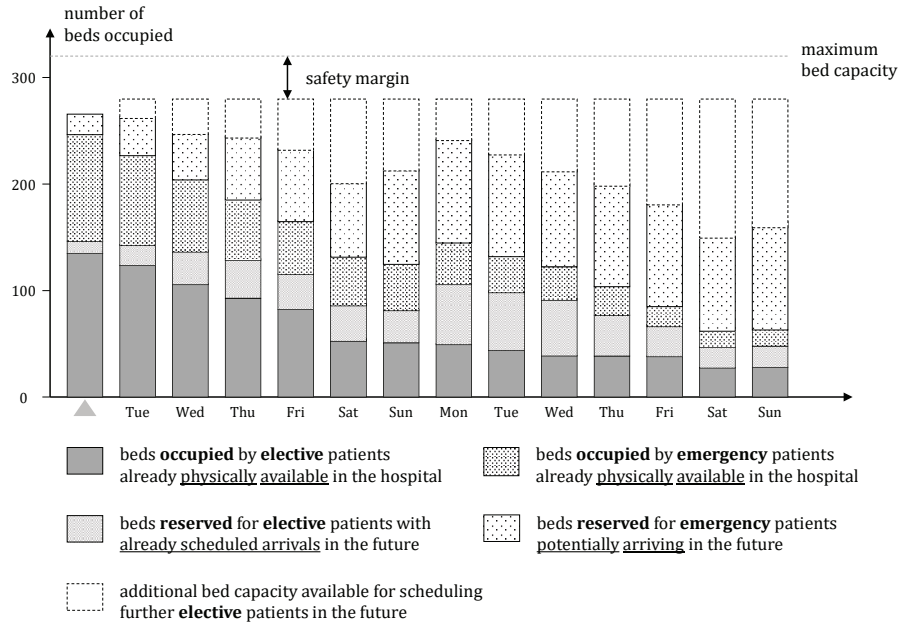


Figure 4.1: Schematic example of a typical planning situation in a large hospital serving elective and emergency inpatients

these four patient classes, respectively, may then be used to schedule additional elective patient arrivals as needed. Please note, patients leaving the hospital on a respective day are already excluded from the bars from a particular day. Newly incoming patients, however, are included in the bars representing the required bed capacity for elective and emergency patients of that day, respectively.

In addition, due to uncertainty regarding the anticipated number of emergency patients as well as LOS changes, a safety margin of beds is established. This is illustrated on the top of all bars in Figure 4.1. The safety margin lowers the available capacity for scheduling elective patients below the maximum possible bed capacity to avoid potential shortages of beds.

Scheduling patients for elective inpatient treatment is usually done a couple of days or even weeks in advance and typically cannot be adjusted at short

notice. This is because elective patients have to prepare for their hospital stay well in advance, e.g., plan and schedule transportation, make necessary arrangements at work and/or at home, or simply have to adhere to certain dietary requirements from their physician in the days leading up to a surgery. In addition, scheduling elective patient arrivals is also dependent on master surgery schedules for patients who require surgery. Master surgery schedules as well as staff rosters and staff scheduling are typically fixed weeks in advance which in turn additionally limits the possibilities for rescheduling patients at short notice (see for example Beliën and Demeulemeester (2007), Bilgin et al. (2012) and Gross et al. (2017)). Finally, emergency patients can typically not be deferred to other hospitals once they have been admitted, i.e., once treatment has started.

It is therefore important to distinguish between PAS and PBA (see Figure 4.2). In PAS elective patients need to be scheduled such that the overall ward utilization is balanced and overflow situations are minimized. In the second step, i.e., the PBA, elective and emergency inpatients need to be assigned actual physical rooms and beds upon entering the hospital. In principle, the PBA problem can be viewed as a downstream decision problem with regard to the PAS problem. For the PAS problem it is not necessary to know which bed exactly will be held available for a certain patient as long as it is guaranteed to a certain extent that a bed will be available.

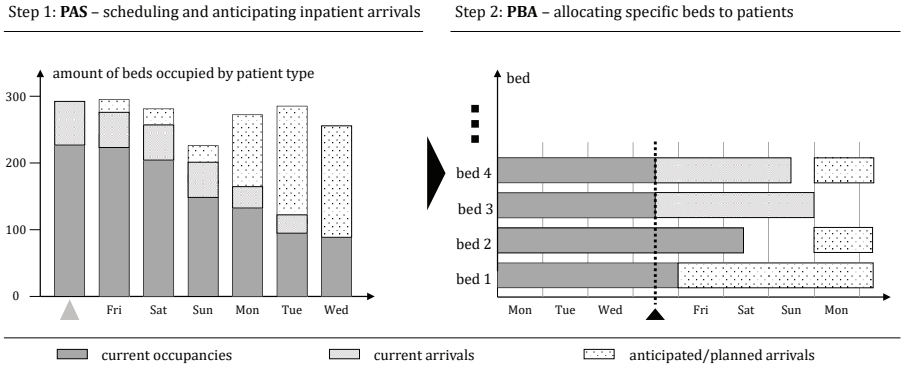


Figure 4.2: Schematic overview of the difference between patient scheduling and patient-bed allocation

The crucial question in PBA is to determine when the actual allocation takes place and whether or not it should be possible to reserve a specific bed for a specific patient in advance prior to their stay. In hospitals equipped with a large number of beds, however, it oftentimes happens that allocation plans made at the beginning of a specific day are obsolete shorty after, due to changes in lengths of stay, no-shows, sudden complications during surgery or treatment, or simply due to emergency arrivals. Thus, any planning system which fixes patient bed allocations several days in advance will inevitably produce allocations that will almost certainly become outdated or even infeasible. Instead, a PBA system should be able to produce a viable allocation for each patient directly when the patient physically needs to occupy his or her room and bed.

In addition, many large hospitals that need to cover all major disciplines exhibit high emergency arrival rates which lead to a higher volatility and uncertainty regarding future occupancy levels. Overflow situations are an inevitable consequence of tight capacities and uncertain demand. In such cases, inpatients need to be assigned to overflow areas such as hallways, emergency- or treatment-rooms, or to other wards outside their dedicated ward space. Staying in such intermediate areas is unpleasant for patients and will always entail additional work for nursing staff and doctors alike, as they typically will not be able to offer the same level of medical assistance. However, the overall LOS of a patient mostly stays the same as necessary surgical procedures and medication treatment will still take place even if a patient is not within his designated ward space. Nevertheless, a bed planner will always try to move patients out of overflow areas whenever the situation allows it to avoid the above-mentioned drawbacks.

As a result of the situation just described, the PBA problem has to be solved several hundred times a day. For each of these planning instances, anticipated future emergency arrivals as well as already scheduled elective inpatients have to be considered. To give an example, a hospital comprising 500 beds and an average LOS of 3 or 4 days requires at least 330 or 250

reruns of the PBA system per day, respectively, i.e., each time a new arrival or departure becomes known to the system.

Objectives In general, patients want their stay to be as pleasant as possible while receiving top-level medical care. This means that patients want to have a room within a designated ward space that caters to their medical needs while avoiding unnecessary room transfers and/or having to wait in an overflow area. In addition, patients want to have pleasant roommates they can get along with and relate to in case they have to share a room. Age difference is a very good indicator for how good patients get along with each other when sharing rooms, especially for longterm stays. This hypothesis was verified by numerous interviews with nursing staff and doctors conducted at our case hospital. Therefore, it is desirable to combine patients of a similar age who have similar illnesses in terms of their specific medical conditions and severity thereof.

As opposed to emergency patients, elective patients are less likely to accept that a room and bed within their respective department is not "reserved" for them upon arrival at the hospital. Emergency patients on the other hand are more willing to accept having to temporarily stay in dedicated overflow areas. In other words, elective patients should in general be preferred when allocating patients to beds during overflow situations. If staying in an overflow area does become necessary, patients wish to be transferred to a "regular room" as soon as possible. In general, it should be noted that elective and emergency patients get the same treatments and the same amount of medical care. The above-described focus on elective patients with regard to patient satisfaction is mainly due to the fact that elective patients will change hospitals for their surgery or treatment if their subjective opinion of a hospital suffers, which would be detrimental to any hospital's reputation.

Doctors are typically bound to a specific department, i.e., a specific medical specialty. In order to facilitate doing rounds and patient visits, it is essential to minimize walking distances for doctors.

One of the main issues when managing patient-bed allocations with regard to nursing staff is creating a balanced workload. This is especially important as nurses are typically dedicated to specific wards in well-coordinated teams, which are used to working with each other and therefore cannot easily be transferred to other wards.

Constraints When trying to optimize PBA, the following hard constraints are typically taken into account. First, non-ICU female and male inpatients are not allowed to be allocated to the same room. Second, certain medical conditions require patients to be in rooms which are equipped with the necessary infrastructure, e.g., telemetry for certain cardiology patients. Third, it may be the case that a patient or several patients need to be isolated from other patients during their stay due to medical reasons. Finally, non-medically induced room transfers are not allowed, meaning that allocations of patients who already physically occupy rooms in their designated department are treated as unchangeable. This is because every physical room transfer entails significant additional work for hospital personnel (e.g., cleaning and sanitizing rooms, moving beds, reorganizing tasks) as well as unnecessary discomfort for the patient. In this context, the only exceptions are transfers due to medical reasons (e.g., transfers to and from the ICU, which may be modeled as separate patient arrivals and discharges).

4.2.2 Related literature and open research questions

Related Literature Scheduling elective inpatients for surgery or treatment such that utilization of bed capacity is optimized has been thoroughly investigated in the literature. For example, Beliën and Demeulemeester

(2007) optimize bed capacity utilization by incorporating the LOS of surgical patients into master surgery schedules in order to balance bed capacity utilization over time. A similar approach has been developed by Fügener et al. (2014) who investigate the effects of scheduling surgery patients on several downstream resources such as the ICU or general ward capacities. Gartner and Kolisch (2014) further investigate scheduling procedures for elective patients such that the contribution margin per patient as well as the utilization of hospital resources such as beds are optimized.

The PBA problem has been introduced by Demeester et al. (2010). Note, that Demeester et al. (2010) define the PBA problem as "patient admission scheduling problem". However, they consider and solve the PBA problem as defined in Section 4.2.1. Demeester et al. (2010) suggest a decision support system that assigns incoming patients to beds. They consider a situation in which a hospital is initially empty and all future patient arrivals within a given time horizon are known as well as their respective parameters, i.e., actual LOS, gender, department adherence, individual infrastructural needs and so forth. In their model, every patient has to be assigned to a room such that an overall cost function based on violating patient-specific requirements and objectives is minimized. The formulated cost function acknowledges gender-specific room allocation, assignment of patients to departments suited for their age, availability of relevant infrastructure, adherence to medical isolation, patient-specific room type preferences (e.g., single or double room) and patient transfers. Based on this cost function patients are assigned to available rooms of a certain type while taking predefined admission and discharge dates of each patient into account. Demeester et al. (2010) neglect nurse- and doctor-specific objectives and do not distinguish between emergency and elective patients. In addition, they assume a static offline planning situation in which all given patients are assigned to the available rooms. An overflow buffer is not considered. Therefore, it has to be ensured in advance that a given data set allows for a feasible assignment of all patients to the limited number of rooms. Demeester et al. (2010) solve the assignment problem using a tabu search algorithm.
Several authors have contributed to the problem of operational bed allo-

cation either by providing alternative and/or improved heuristic solution approaches for the problem defined by Demeester et al. and/or by adding certain aspects to the problem.

Ceschia and Schaerf (2011) build on the model, solution approach, and data sets provided by Demeester et al. (2010) by introducing new neighborhood search strategies. They further propose a relaxation procedure to provide lower bounds and introduce a simple dynamic version of the planning problem. Subsequently, the authors expanded on their work and introduced a more sophisticated heuristic solution approach involving simulated annealing, incorporated emergency patient arrivals (see Ceschia and Schaerf (2012)), and most recently included operating room utilization (see Ceschia and Schaerf (2016)). Additionally, Ceschia and Schaerf (2016) allow admission delays while penalizing delays that happen close to the originally planned admission date but do not consider overflow per se.

Bilgin et al. (2012) build on the work of Demeester et al. (2010) by investigating a hyper-heuristic approach to the PBA problem which focuses on optimizing the trade-off between run-time and solution quality. A different solution approach similarly aimed at finding a faster solution approach was proposed by Range et al. (2014) who use a column generation approach for solving the PBA problem. Vancroonenburg et al. (2016) propose to divide the PBA problem into two IP models which assign current patients to beds and reserve beds for future patient arrivals, respectively. A further approach to solving the PBA problem worth noting was presented by Schmidt et al. (2013), which in contrast to the afore-mentioned approaches, focuses on assigning patients to bed contingents rather than individual beds, i.e., they neglect room- and bed-specific characteristics, while respecting a given set of patient preferences, respectively.

Open research and contribution to the literature In our joint project with a large German hospital we identified a variety of additional aspects which to the best of our knowledge have not been dealt with in the literature

currently available regarding the PBA problem. We therefore suggest a more comprehensive decision support model and a specialized solution approach that overcomes actual planning shortages. The new modeling and solution approach respects diverse interests of patients, nurses, doctors and hospital management while simultaneously considering several hard constraints when assigning patients to rooms and beds, i.e., medical, gender as well as capacity constraints. In addition, we explicitly distinguish between emergency and elective patients and consider their specific needs and requirements. Furthermore, we deal with ad-hoc overflow situations in which it is not possible to simply reschedule or defer patients. We assume a dynamic online planning situation in which the PBA problem needs to be solved several hundred times a day, i.e., at each point in time an inpatient gets admitted or discharged or when any other change in the system merits moving patients from an overflow area to a regular bed. In addition, the developed approach is based on real time data that also anticipates future developments, such that the decision support system can provide reliable online recommendations for patient-bed allocations. Last but not least we prove the general applicability of the approach suggested in hospital practice using data sets from the literature as well as actual clinic data from our case study hospital.

4.3 Modeling and solution approach

In the present section we develop a decision support model and a greedy look-ahead heuristic (GLA heuristic) to assign elective and emergency inpatients to beds. The model and the solution approach is designed to be solved every time a change in the underlying parameters of the system may lead to the physical allocation of a newly arrived patient or a patient waiting in an overflow area to a regular bed. This may lead to several hundred reruns of the designed procedure per day.

4.3.1 Model development

The deterministic model maximizes a utility function which quantifies the trade-off between patient-specific, doctor-specific, as well as nurse-specific objectives, while simultaneously considering medical, gender as well as capacity constraints when assigning patients to rooms and beds. In addition, the model allows to assign patients to an overflow area if regular beds are not available during the first or — as the case may be — for up to all days of their designated stay. Table 4.1 summarizes the sets, parameters and variables used when formulating the model.

Sets	
B	set of beds which are scheduled to be vacated within the planning horizon of $\|T\|$ days, $B = \{1, 2, ..., b, ..., \|B\|\}$
D	set of departments, $D = \{1, 2, ..., d, ..., \|D\|\}$
P	set of patients who require a bed at some point in time within the planning horizon of $\|T\|$ days including patients already waiting in the overflow area, $P = \{1, 2, ..., p, ..., \|P\|\}$
R	set of rooms which have at least one available bed, $R = \{1, 2, ..., r, ..., \|R\|\}$
T	set of days within the planning horizon, $T = \{1, 2, ..., t, ..., \|T\|\}$
W	set of wards which have at least one available bed, $W = \{1, 2, ..., w, ..., \|W\|\}$
Parameters	
$\alpha, \beta, \gamma, \delta$	weighting factors for patient- (α and β), doctor- and nurse-related utilities, respectively
Ξ_p	weighting factor that allows to distinguish between patient types, e.g., elective and emergency patients
A_p	age of patient p
$\mathrm{A}_{rt}^{\max}$ $\left(\mathrm{A}_{rt}^{\min}\right)$	$\mathrm{A}_{rt}^{\max}$ $\left(\mathrm{A}_{rt}^{\min}\right)$ is set to the maximum (minimum) age of all patients already physically occupying room r for the night on day t and to 0 (M) if the room is empty
OV_p	utility parameter depending on the time patient p has already spent in the overflow area due to a previous overflow situation

Continued on next page

Table 4.1 – *Continued from previous page*

c_{wt}	additional care capacity for scheduling additional patients $p \in P$ on ward w on day t
C_p	care level required to accommodate patient p
d_{rt}	d_{rt} represents the department of the prior occupants of room r on day t only in case all of them are allocated to the same department and 0 otherwise
D_p	associated department of patient p with $D_p \in D$
e_{bt}	$e_{bt} = 1$ if bed b is located in a room that is initially empty on day t and 0 otherwise
f_{rt}	$f_{rt} = 1$ if room r is initially empty on day t and 0 otherwise
G_p	$G_p = -1$ if patient p is male and $G_p = 1$ if patient p is female
I_p	$I_p = -1$ if patient p requires medical isolation and 1 otherwise
K_{br}	$K_{br} = 1$ if bed b is in room r and 0 otherwise
L_{bw}	$L_{bw} = 1$ if bed b is in ward w and 0 otherwise
M	large integer value
π_{bp}	utility of assigning patient p to bed b based on overflow and patient type (basic model)
Q_t	relevance of a bed allocation for a patient on day t as anticipated/planned; Q_t approximated as $Q_t = (1-q)^t$ with discounting parameter $q \in]0;1[$
s_{bpt}	$s_{bpt} = 1$ in case bed b is available for patient p on day t of his stay in the hospital and 0 otherwise ("availability" further considers gender, infrastructural, and medical isolation constraints based on pre-occupancies in the room of bed b)
Decision variable	
x_{bp}	$x_{bp} = 1$ if patient p is assigned to bed b and 0 otherwise
Auxiliary variables	
$a_{rt}^{\max}$ $\left(a_{rt}^{\min}\right)$	$a_{rt}^{\max}$ $\left(a_{rt}^{\min}\right)$ is the maximum (minimum) age of all patients p assigned to room r on day t

Continued on next page

Table 4.1 – *Continued from previous page*

o^+_{wt}	o^+_{wt} denotes the additional accumulated care level surpassing a predefined threshold for a given ward w on day t
y_{rt}	$y_{rt} = 1$ if all patients assigned to an empty room r on day t are from the same department and 0 otherwise
z_{rt}	$z_{rt} = 1$ if all patients assigned to a partially occupied room r are from the same department as the patients already occupying room r and 0 otherwise

Table 4.1: Notation

We formulate the objective function as a multi-objective utility maximization function to accommodate the trade-offs between the diverse interests of patients, nurses and doctors that exist when allocating patients to beds. The objective function (4.1) is formulated as follows:

$$\max \ \Pi = \alpha \ f_{\text{basic}}(x_{bp}) - \beta \ f_{\text{patient}}(x_{bp}) + \gamma \ f_{\text{doctor}}(x_{bp}) - \delta \ f_{\text{nurse}}(x_{bp}) \tag{4.1}$$

Equation (4.1) consists of four terms that represent (I) basic patient-specific objectives, (II) extended patient-specific objectives, (III) doctor-specific objectives and finally (IV) nurse-specific objectives. In the following, we will gradually develop the four parts. The four partial objectives are weighted by the factors α, β, γ, and δ. These weighting factors are used to control the influence of the individual objectives on the overall solution. They are derived from managerial decisions. All four objective values depend on the assignment variable x_{bp} which equals to 1 if patient p is allocated to bed b and 0 otherwise.

(I) Basic patient-specific objectives and constraints The first term quantifies the patient-type-specific objectives:

$$f_{\text{basic}}(x_{bp}) = \sum_{b \in B} \sum_{p \in P} \pi_{bp} x_{bp} \tag{4.2}$$

Parameter π_{bp} denotes the patient-type-specific "utility" of assigning patient p to bed b. It depends solely on information known prior to updating the bed allocation planning. Thus, parameter π_{bp} is not influenced by other assignments of patients $p \in P$ to beds $b \in B$ during a specific planning instant. As room transfers are not allowed, every assignment of a patient p to a bed b, i.e., $x_{bp} = 1$ generates a utility of π_{bp}, which accounts for the days that patient p actually spends in bed b within the planning horizon T. For a given patient p and a given bed b the utility value is quantified as follows:

$$\pi_{bp} = \mathrm{OV}_p + \Xi_p \sum_{t \in T} \mathrm{s}_{bpt} \mathrm{Q}_t \tag{4.3}$$

Here, OV_p represents a predetermined utility value which depends on the time a patient p has already spent in the overflow area in the past. This "overflow bonus" is only awarded to patients who are already waiting in the overflow area at the time the decision model is solved. This is done to ensure that patients who are already in the overflow area do not risk staying there for the entirety of their stay. In other words, the set of patients P includes not only current and future planned and anticipated emergency patient arrivals but also patients that are currently waiting in the overflow area. Patients already waiting in the overflow area will be preferred to otherwise similar patients who have just arrived in the hospital as a result of the additional utility value OV_p.

The second part of Equation (4.3) rewards the actual time that a patient p spends in one of the beds $b \in B$. To this end, the predetermined parameter s_{bpt} is introduced, which is preset to 1 in case bed b is available for patient p on day t and 0 otherwise. As non-medical room transfers are not allowed, s_{bpt} can be determined entirely during preprocessing and is used to reflect not only bed availability but also bed compatibility by incorporating gender constraints, infrastructural constraints, as well as medical isolation constraints for each possible patient-bed combination. The advantage of summing up s_{bpt} over $t \in T$ can be seen in that s_{bpt} may be determined entirely during preprocessing. Figure (4.3) shows an example illustrating how parameter s_{bpt} is set to 0 or 1. The example given considers 4 rooms, i.e., room I with beds 1 and 2, room II with beds 3 and 4, and so forth. Here, there are multiple options for allocating female patient 1 to a bed. This patient arrives at the hospital on day 3 and is scheduled to be discharged on day 8, thus having an anticipated LOS of 5 days. An allocation to bed 1, for example, would imply an initial stay of three days within the overflow area before moving to bed 1 for the remaining two days. Accordingly, an allocation to bed 1 would have a lower utility than an allocation to bed 4, for example, as the first three days spent in the overflow area do not create any additional benefit. Beds 5 and 6, for example, are not allowed to be used as this room is not equipped with essential medical infrastructure specifically required for treating patient 1. Beds 7 and 8, however, are both currently occupied by female patients requiring medical isolation from non-quarantined patients for the duration of their stay, hence forcing patient 1 to spend one day within the overflow area before moving into either of these beds, should she be allocated to one of them.

It is important to note, that s_{bpt} does not define the LOS of patient p. This is because treatment of patients (e.g., surgery, medication) typically starts once the patient arrives at the hospital, regardless of where in the hospital their bed is physically located.

In addition, x_{bp} defines not only the bed b that patient p is allocated to, but also the time patient p has to spend in the overflow area depending

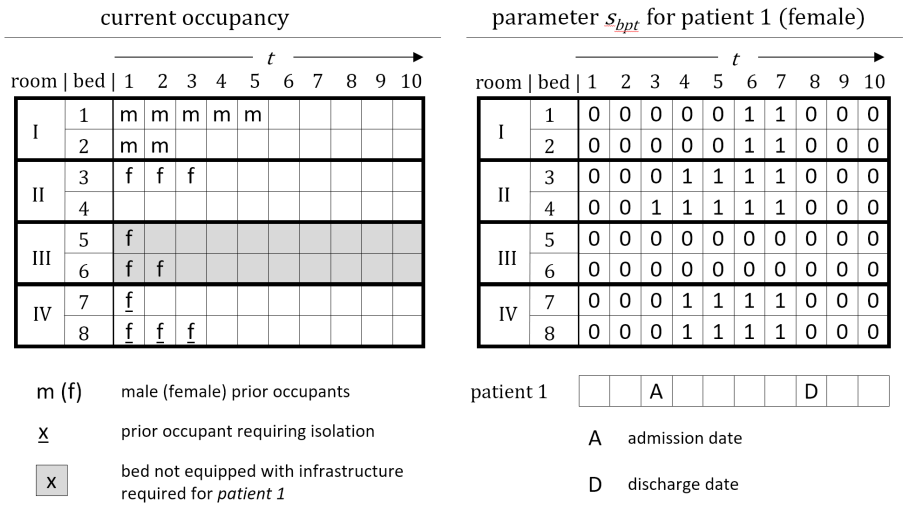


Figure 4.3: Example for quantifying s_{bpt}

on the current occupancy situation at the time the planning is updated. In addition, spending time in the overflow area does not affect the overall LOS. The parameter Ξ_p is a factor that allows to distinguish between patient types, i.e., elective patients, emergency patients, or patients with special infrastructural requirements. This factor may, for example, be used to ensure that elective patients are more likely to be assigned to a bed within their target ward upon arrival than emergency patients, or to ensure that patients with special infrastructural needs are preferred. For example, patients returning from the ICU could be attributed an even higher value such that it is highly unlikely for them to be moved to an overflow area.

Finally, $\sum_{t \in T} s_{bpt} Q_t$ incorporates the time a patient is assigned to a regular bed during his/her LOS. Q_t is a parameter that reflects the relevance of a bed allocation for a patient on day t as anticipated/planned where Q_t is decreasing with increasing t. Thus, otherwise similarly evaluated patients contribute to the overall objective function with a higher utility if they require a bed earlier in the planning horizon considered. This modeling approach anticipates the possibility of reassigning later arriving patients

to other beds at planning instants in the future. Due to uncertainties it is quite reasonable that a patient, who is planned to arrive far in the future, will be reassigned to another bed at later planning periods, which may then even lead to a higher overall utility value for that patient. Possible uncertainties are related to LOS, emergency arrivals, treatment progression, no-shows and so forth. The decreasing parameter Q_t is approximated as follows, assuming $\mathrm{q} \in]0;1[$:

$$\mathrm{Q}_t = (1-\mathrm{q})^t \tag{4.4}$$

In the following, the basic set of hard constraints is listed which have to be adhered to regardless of how the individual parts of the objective function are actually weighted.

$$\sum_{b \in B} x_{bp} \leq 1 \qquad \forall p \in P \tag{4.5}$$

$$\sum_{p \in P} \mathrm{s}_{bpt} x_{bp} \leq 1 \qquad \forall b \in B; t \in T \tag{4.6}$$

$$\sum_{t \in T} \mathrm{s}_{bpt} \geq x_{bp} \qquad \forall b \in B; p \in P \tag{4.7}$$

$$\mathrm{K}_{br}\mathrm{e}_{bt}\mathrm{G}_p\mathrm{s}_{bpt}x_{bp} - \mathrm{K}_{lr}\mathrm{e}_{lt}\mathrm{G}_h\mathrm{s}_{lht}x_{lh} \geq -1 \quad \forall b,l \in B; p,h \in P; r \in R; t \in T \tag{4.8}$$

$$\mathrm{K}_{br}\mathrm{e}_{bt}\mathrm{I}_{p}\mathrm{s}_{bpt}x_{bp} - \mathrm{K}_{lr}\mathrm{e}_{lt}\mathrm{I}_{h}\mathrm{s}_{lht}x_{lh} \geq -1 \qquad \forall b, l \in B; p, h \in P; r \in R; t \in T \tag{4.9}$$

$$x_{bp} \in \{0, 1\} \qquad \forall b \in B; p \in P \tag{4.10}$$

Equation (4.5) prevents double-booking by ensuring that each patient is assigned to no more than one bed. Please note, a patient receives no bed assignment if he/she entirely stays in the overflow area during his/her scheduled LOS. In addition, Equation (4.6) prevents overbooking, such that no two patients are allocated to the same bed on the same day. Equation (4.7) ensures that a patient p can only be assigned to a bed b, i.e., $x_{bp} = 1$ if bed b is at least available for this patient on one day of the planning horizon, i.e., $s_{bpt} = 1$ for at least one $t \in T$.

Furthermore, Equation (4.8) in combination with s_{bpt} ensures that there are no mixed male and female rooms on any given day t. Here, G_p is set to -1 if patient p is male and to 1 if patient p is female. In particular, Equation 4.8 compares all patients $p \in P$ which may be allocated to room r on day t. In case a male patient is mixed with a female patient, the equation would not be satisfied as it would then read $-1 - 1 \geq -1$, i.e., $-2 \geq -1$. In addition, most rooms are already preoccupied on specific days, such that only male or female patients are additionally allowed, respectively. As pointed out above, this prior occupancy is integrated into s_{btp}. Prior occupancies are reflected in s_{bpt} such that $\mathrm{s}_{bpt} = 0$ for a female patient p in case a bed b is located in a room which is still occupied by at least one male patient on day t and vice versa (see Figure (4.3) for an example). The parameter e_{bt} is set to 1 if bed b is located in a room that does not have any current occupants (i.e., which is empty at the time the planning is updated) on day

t and 0 otherwise. Finally, K_{br} connects beds to rooms and is set to 1 if bed b is located in room r and 0 otherwise.

Using a similar approach, equation (4.9) in combination with s_{bpt} ensures that medical isolation requirements are respected. Specifically, patients that need to be isolated due to infectious diseases, for example, may only be put into empty rooms or into rooms with patients that suffer from the same condition. Here, I_p is set to -1 if patient p requires medical isolation and 1 otherwise.

(II) Further patient-specific objectives and constraints The second term of the objective function (4.1) is used to model the preferences of the patients. This part of the objective function tries to minimize the age differences within rooms since it is desirable to combine patients of a similar age as it is more likely for them to share common interests. In addition, they potentially share similar illnesses when associated to the same medical department. Numerous interviews at our case hospital verify this approach. The second term of the objective function is then denoted as follows:

$$f_{\text{patient}}(x_{bp}) = \sum_{r \in R} \sum_{t \in T} (a_{rt}^{\max} - a_{rt}^{\min}) \tag{4.11}$$

Here, $a_{rt}^{\max}$ ($a_{rt}^{\min}$) denotes the maximum (minimum) age of all patients which during run-time of the model are going to be assigned to room r on day t. As such, both auxiliary variables $a_{rt}^{\max}$ and $a_{rt}^{\min}$ are dependent on the overall decision variable x_{bp}. The following constraints are used to link these auxiliary variables to x_{bp}:

$$a_{rt}^{\max} \geq \mathrm{A}_{rt}^{\max} \qquad \forall r \in R; t \in T \tag{4.12}$$

$$a_{rt}^{\max} \geq \mathrm{K}_{br}\mathrm{A}_{p}\mathrm{s}_{bpt}x_{bp} \qquad \forall b \in B; p \in P; r \in R; t \in T \quad (4.13)$$

$$a_{rt}^{\min} \leq \mathrm{A}_{rt}^{\min} \qquad \forall r \in R; t \in T \quad (4.14)$$

$$a_{rt}^{\min} \leq \sum_{b \in B} \sum_{p \in P} \mathrm{A}_{rt}^{\min}\mathrm{K}_{br}\mathrm{s}_{bpt}x_{bp} \qquad \forall r \in R; t \in T \quad (4.15)$$

$$a_{rt}^{\min} \leq \mathrm{K}_{br}\mathrm{A}_{p}\mathrm{s}_{bpt}x_{bp} + \mathrm{A}_{rt}^{\min}(1 - x_{bp}) \quad \forall b \in B; p \in P; r \in R; t \in T \quad (4.16)$$

Here, $\mathrm{A}_{rt}^{\max}$ is set to the current maximum age of all patients already occupying room r on day t and to 0 in case room r is empty on day t. As it is solely dependent on prior occupancy, $\mathrm{A}_{rt}^{\max}$ is determined entirely during preprocessing and is not affected by x_{bp}.

With the same logic, $\mathrm{A}_{rt}^{\min}$ is set to the minimum age of all patients already occupying room r on day t and to a large integer value, e.g., 150, the maximum age of any possible patient, in case there are no prior occupants in room r on day t. Thus, Equations (4.12) and (4.13) ensure that the auxiliary variable $a_{rt}^{\max}$ reflects the maximum age of prior occupants and newly allocated patients in a room r on day t. Likewise, Equations (4.14) to (4.16) ensure the same for $a_{rt}^{\min}$ while also making sure that $a_{rt}^{\min}$ equals $a_{rt}^{\max}$ in the case room r is only occupied by one person or completely empty on day t.

(III) Doctor-specific objectives and constraints The third term of the objective function (4.1) rewards assigning patients of the same department to identical rooms. Medical rounds for doctors are easier when having several patients they are responsible for in the same room. In addition, walking distances are reduced. The third term of the objective function is then formulated as follows:

$$f_{\text{doctor}}(x_{bp}) = \sum_{r \in R} \sum_{t \in T} \mathrm{f}_{rt} y_{rt} + \sum_{r \in R} \sum_{t \in T} (1 - \mathrm{f}_{rt}) z_{rt} \quad (4.17)$$

As before, an additional set of constraints is required to establish the link between the decision variable x_{bp} and Equation (4.17):

$$\mathrm{K}_{br}\mathrm{D}_{p}\mathrm{s}_{bpt}x_{bp} - \mathrm{K}_{lr}\mathrm{D}_{h}\mathrm{s}_{lht}x_{lh} \geq -M(1-y_{rt}) \qquad \forall b,l \in B; p,h \in P; r \in R; t \in T \tag{4.18}$$

$$\sum_{p\in P}\sum_{b\in B}\mathrm{K}_{br}\mathrm{s}_{bpt}x_{bp} \geq y_{rt} \qquad \forall r \in R; t \in T \tag{4.19}$$

$$\mathrm{K}_{br}\mathrm{D}_{p}\mathrm{s}_{bpt}x_{bp} - \mathrm{d}_{rt} \leq M(1-z_{rt}) \qquad \forall b \in B; p \in P; r \in R; t \in T \tag{4.20}$$

$$\mathrm{d}_{rt} - \mathrm{K}_{br}\mathrm{D}_{p}\mathrm{s}_{bpt}x_{bp} \leq M(1-z_{rt}) \qquad \forall b \in B; p \in P; r \in R; t \in T \tag{4.21}$$

$$\sum_{p\in P}\sum_{b\in B}\mathrm{K}_{br}\mathrm{s}_{bpt}x_{bp} \geq z_{rt} \qquad \forall r \in R; t \in T \tag{4.22}$$

$$y_{rt} \in \{0,1\} \qquad \forall r \in R; t \in T \tag{4.23}$$

$$z_{rt} \in \{0,1\} \qquad \forall r \in R; t \in T \tag{4.24}$$

Here, D_p is an integer parameter that depicts the department that corresponds to the medical condition of patient p. In addition, d_{rt} depicts the department of all prior occupants of room r on day t only if all prior occupants are from the same department and is set to 0 otherwise. As such, both D_p and d_{rt} are determined entirely during preprocessing. The parameter f_{rt} is 1 in case room r does not have any prior occupants on day t and 0 otherwise. Accordingly, the auxiliary variable y_{rt} is set to 1, if all patients assigned to an empty room r on day t are from the same department which is achieved by Equations (4.18) and (4.19). An additional auxiliary variable z_{rt} is used in case a room r is already preoccupied on day t and is set to 1 only if all patients assigned to room r as well as the patients in room r are already from the same department. This is achieved with Equations (4.20) to (4.22).

(IV) Nurse-specific objectives and constraints Finally, the fourth term of the objective function (4.1) is used to balance workload for nursing staff (see Section 4.2 for details) and is quantified as follows:

$$f_{\text{nurse}}(x_{bp}) = \sum_{w \in W} \sum_{t \in T} o^{+}_{wt} \tag{4.25}$$

In particular, exceeding a predefined care capacity for nursing staff assigned to ward w on day t is penalized. To this end, the following additional set of constraints is required:

$$\sum_{b \in B} \sum_{p \in P} \mathrm{L}_{bw} \mathrm{C}_{p} \mathrm{s}_{bpt} x_{bp} \leq \mathrm{c}_{wt} + o^{+}_{wt} \qquad \forall t \in T; w \in W \tag{4.26}$$

$$o^{+}_{wt} \geq 0 \qquad \forall t \in T; w \in W \tag{4.27}$$

Parameter C_p quantifies the level of care required for patient p. This represents the effort and resources that go into taking care of a particular patient. In addition, the available number of nursing staff and thus, workforce per ward w and day t is predetermined due to shift schedules, staff rosters, and so forth. Thus, the parameter c_{wt} represents the additional care capacity of a given ward w on day t, i.e., the capacity to take in additional patients $p \in P$ requiring C_p units of care, respectively. For instance, assume $c_{wt} = 6$ for a given ward w and a given day t. This would then mean that ward w could additionally take up 2 patients with a care level $C_p = 3$ before overloading the nursing staff of that ward on that day, for example. Nursing staff typically cannot be moved from ward to ward on an ad-hoc basis. This means that having patients in a first ward that are very easy to handle cannot balance out a second ward filled with a very labor-intensive patient clientele. Thus, the auxiliary variable o^+_{wt} is introduced which denotes the additional accumulated care level surpassing the predefined care capacity threshold for a given ward w on day t. Equations (4.26) and (4.27) are used to link x_{bp} to o^+_{wt}.

4.3.2 Greedy look-ahead heuristic

An efficient bed allocation support system needs to be able to give a bed planner online recommendations for patient bed allocations within seconds when prompted. This is due to real-life planning situations in large hospitals requiring highly flexible planning systems which are able to adapt to ad-hoc changes in real time. However, solving the model by Gurobi's MIP solver requires more than 12 hours for relevant problem instances (see Section 4.4 for details). Likewise, other approaches followed in the literature (see for example Demeester et al. (2010); Ceschia and Schaerf (2011)) also had to resort to using heuristic approaches for the same reasons. We therefore develop a novel greedy look-ahead heuristic (GLA heuristic) which bases on the general idea of Atkinson (1994) by sequentially assigning the most

utility-attractive patient to his or her most beneficial bed while anticipating potential room allocations still to be made in futher steps of the algorithm. Table 4.2 summarizes the additional notation required to formulate the GLA heuristic.

U_{bp}	partial utility that an allocation of patient p to bed b may add to the overall utility Π, $p \in P$, $b \in B$
U_p^{argmax}	index value of the bed that adds the maximum partial utility to the overall utility Π when patient p, $p \in P$ will be allocated to this bed
$U_p^{\max}$	maximum partial utility that an allocation of patient p may add to the overall utility Π, $p \in P$

Table 4.2: Expanded notation for the GLA heuristic

The basic premise of the GLA heuristic is based on a greedy algorithm when assigning patients to beds which approximates assignments of patients to beds that may be realized in later stages of the algorithm. To this end, a utility matrix U_{bp} is used which, upon initiation of the PBA-algorithm, is prefilled with the partial utilities that a respective allocation of patient p to bed b would add to the overall utility Π of the objective function (4.1). It should again be noted in this context, that the set of patients P as well as the set of beds B includes not just current but also future patient arrivals and bed availabilities, respectively, and as such every value of U_{bp} implicitly includes time already spent in and time to be spent in the overflow area as well as uncertainty regarding future arrival and discharge dates. Should a specific bed b not be available at all for patient p at any time of their planned stay, the value U_{bp} is set to zero.

Upon initiation of the GLA heuristic, x_{bp} is set to 0 for all $b \in B$ and $p \in P$. As described above, the initial values for U_{bp} are calculated for every $b \in B$ and $p \in P$. Subsequently, the highest value in U_{bp} is identified and x_{bp} is set to 1 correspondingly, i.e., patient p is allocated to bed b. Finally, all elements in U_{bp} that are affected by any allocation are updated

before the next patient is allocated. To streamline the computations, only the vectors $U_p^{\max}$ and U_p^{argmax} are calculated. $U_p^{\max}$ contains the maximum partial utility that an allocation of patient p may add to the overall utility Π and U_p^{argmax} reveals the index value of the corresponding bed b. If necessary the values $U_p^{\max}$ and U_p^{argmax} are also updated after every allocation. This way, the PBA-algorithm only has to compare $|P|$ values instead of $|P| \times |B|$ values.

Figure 4.4 illustrates the first steps of the GLA heuristic. Step 1 of Iteration I shows the initial utility matrix U_{bp} as well as the initial corresponding values for $U_p^{\max}$ and U_p^{argmax}. The highest value of U_{bp} then determines the first allocation, i.e., x_{62} is set to 1. This initial allocation of patient $p = 2$ to bed $b = 6$ then has an effect on a series of potential allocation combinations x_{bp} of the remaining patients P and beds B. Therefore, in Step 2 of Iteration I the utility matrix U_{bp} is updated and if necessary the variables $U_p^{\max}$ and U_p^{argmax} are redetermined. In the example shown in Figure 4.4, the values marked with black boxes were updated. Iteration II is then substantially equivalent and subsequent to Iteration I. Algorithm 4.1 summarizes the sequential, procedural program flow.

Algorithm 4.1 GLA heuristic

Require: P, B
Ensure: patient-bed allocations x_{bp}
1: $U_{bp} \leftarrow$ calculatePatientBedMatrix(P, B)
2: $U_p^{\max} \leftarrow \max(U_{bp})$
3: $U_p^{\text{argmax}} \leftarrow \text{argmax}(U_{bp})$
4: **while** $(\max(U_{bp}) \neq 0)$ **do**
5: $\quad p \leftarrow \text{argmax}(U_p^{\max})$
6: $\quad b \leftarrow U_p^{\max}[p]$
7: $\quad x_{bp} \leftarrow 1$
8: $\quad U_{bp} \leftarrow$ updatePatientBedMatrix(p, b, U_{bp}, P, B)
9: **end while**
10: printPatientBedAllocations(x_{bp})

Allocating patients to rooms that are “still empty” at the time of allocation but will be filled during later iterations, i.e., during runtime of the algorithm, is approximated as follows. The value B_{bp} for the case that patient p is allocated to bed b in a previously unoccupied room (at this exact point in the GLA heuristic run through) is calculated assuming that any potential

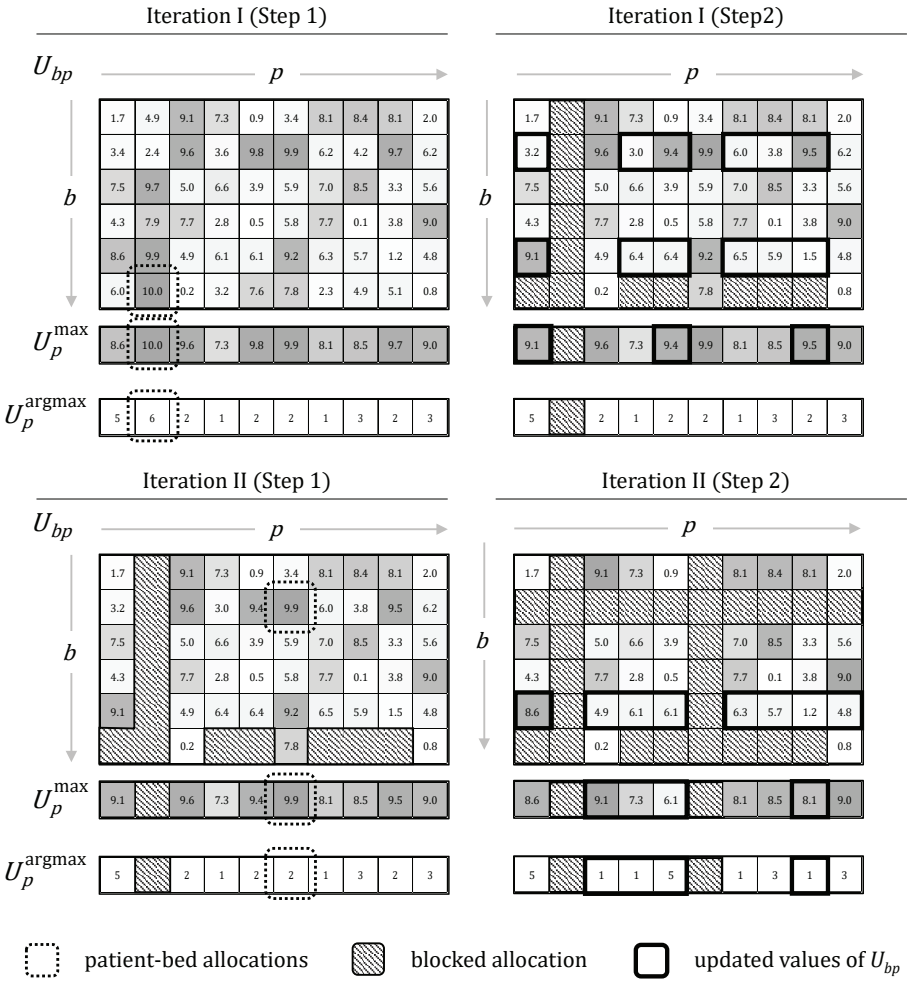


Figure 4.4: Example of the GLA heuristic

future room-mates will not have the same department and will have the largest possible age-difference based on the pool of patients that are still to be allocated during the current run-through of the GLA heuristic. The approach, therefore, "looks ahead" or approximates potential assignments of rooms which will happen at a later stage of the run-through of the algorithm. The procedure avoids that patients are disproportionately assigned to so far empty rooms.

4.4 Numerical study

In this section we present detailed results for our proposed approach. First, the choice of parameters generally used for the numerical studies is stated in Section 4.4.1. In Section 4.4.2 we assess the performance of the GLA heuristic by comparing runtime and solution quality of the GLA heuristic with near-to-optimal solutions obtained by solving our model with Gurobi's MIP solver. In Section 4.4.3, we then solve a case study for a large German hospital. In Section 4.4.4 we analyze the general applicability of our approach by employing different sized problem instances from literature. Section 4.4.5 then further investigates our contribution to literature with regard to patient-specific, doctor-specific, and nurse-specific objectives. All computational steps were carried out in Python 3.6.3 and Gurobi 7.5. All computations were run on a work station equipped with 2 Intel Core E5-2620 processors and 64 GB of RAM.

4.4.1 Parameters

The parameters presented in this Section are used for the following numerical tests. In discussions with nurses, doctors, and hospital management we determined the basic parameters to be used for our case study (see Table 4.3).

Parameters	α	β	γ	δ	Ξ^{el}	Ξ^{em}	Ξ^{an}	Q_t
Values	1	0.01	0.2	0.2	20	19	3	$Q_t = (1-q)^t; q = 0.01$

Table 4.3: Overview of weighting factors used

The main goal was to ensure that elective patients are generally preferred over emergency arrivals to prevent allocating them to the overflow area (see Section 4.2 for details). To achieve this, the weighting factor Ξ_p was set to three distinct values depending on the patient type. Notably, these consist of Ξ^{el} for elective patients, Ξ^{em} for current emergency arrivals, and Ξ^{an} for anticipated emergency arrivals. Here, current emergency arrivals are preferred over anticipated future emergency arrivals. This is due to the fact, that the parameters of recently arrived emergency patients requiring a bed are well known, whereas the relevant parameters of future emergency arrivals have to be anticipated based on historical probability distributions. Finally, the parameters α, β, γ, δ, and Q_t were set such that patient-specific, doctor-specific, and nurse-specific objectives reflect managerial decisions regarding PBA in our case hospital. In our case study, two specific effects regarding uncertainty of future events stood out. First, the no-show probability was higher, the farther in the future a patient arrival was scheduled. This is to be expected, as the time for potential problems or issues to arise is longer. Second, doctors responsible for giving LOS estimates based on their patients' medical conditions tended to be more conservative with these estimates the longer the remaining LOS was. This was mainly due to doctors wanting to "avoid false promises" to patients and the admission scheduling office alike. We approximate these issues by a geometric function $Q_t = (1-q)^t$ wherein q represents the associated discounting factor. In essence, patients requiring a bed earlier within the planning horizon obtain a higher priority than those who require a bed at a later stage.

4.4.2 Performance of the GLA heuristic

To assess the overall solution quality of our GLA heuristic we created different sized problem instances for testing purposes ranging from 24 to 120 beds while using a planning horizon between 1 and 9 days. For each of these problem sizes, we created 20 unique data sets based on the original data obtained from our case hospital, i.e., over a year. In particular, 20 different "snap shots in time" were chosen at random from a 9 month period worth of raw data to provide 20 completely different starting situations or problem instances.
We then compared average run times for each problem size by comparing near to optimal solutions of a Gurobi implementation of our model with results obtained by our GLA heuristic (see Table 4.4). Near to optimal means that we allowed a MIP Gap of up to 1%.

GLA heuristic - average solution time in seconds					
|B|	|T| = 1	|T| = 3	|T| = 5	|T| = 7	|T| = 9
24	0.005	0.007	0.014	0.014	0.095
48	0.012	0.028	0.060	0.106	0.171
72	0.015	0.051	0.165	0.261	0.445
96	0.022	0.073	0.157	0.296	0.551
120	0.044	0.139	0.246	0.585	0.853
Gurobi solution - average solution time in seconds					
|B|	|T| = 1	|T| = 3	|T| = 5	|T| = 7	|T| = 9
24	17	125	740	2124	9360
48	157	3492	19780		
72	1212	16476			
96	2822			stopped after 12 hours	
120	10479				

Table 4.4: Computational time analyses

Table 4.4 gives an overview of the average run times obtained. For the smallest problem size, i.e., |B| = 24, the run time of the Gurobi implementation increases considerably from 17 seconds when only considering a planning horizon of |T| = 1 to over 2.5 hours when using a planning

horizon of $|T| = 9$. A similar increase can be observed when augmenting the amount of beds included in the problem. Hence, run time heavily depends on both planning horizon and the amount of beds considered such that typical problem sizes, e.g., 100 beds and more with a planning horizon of 1 week, cannot be solved within a reasonable time frame. For the purpose of our analyses we stopped the Gurobi solver after 12 hours for each data set. However, when using our GLA heuristic, solution times stayed at under a second even for the largest problem size tested. In addition, the solution times obtained by the heuristic show a significantly lower rate of increase compared to the Gurobi implementation when moving from smaller to larger problem sizes.

Gurobi solution - average MIP Gap					
$\|B\|$	$\|T\| = 1$	$\|T\| = 3$	$\|T\| = 5$	$\|T\| = 7$	$\|T\| = 9$
24	0.46%	0.53%	0.57%	0.72%	0.79%
48	0.80%	0.87%	0.89%	3.59%	6.79%
72	0.84%	0.95%	3.78%	6.82%	7.30%
96	0.96%	6.45%	6.70%	7.54%	8.19%
120	0.98%	6.88%	7.87%	10.60%	15.39%

Table 4.5: Overview of average MIP Gap of the Gurobi implementation

Table 4.5 shows an overview of the average MIP Gaps obtained with the Gurobi implementation. The near-to-optimal solutions shown above the dashed lines were all able to be solved with a MIP Gap of about 1% or less within less than twelve hours (see Table 4.4). For all other values, the Gurobi MIP solver was stopped after 12 hours and the respective solutions and their corresponding MIP Gap at that time were recorded. Here, it can be seen that solving the model in adequate time with a standard MIP program does not seem feasible for typical problem instances in large hospitals.

Table 4.6 shows the average as well as the minimum and maximum solution quality obtained for each problem size. Solution quality is defined as

GLA heuristic - average solution quality compared to Gurobi solution[1]					
|B|	|T| = 1	|T| = 3	|T| = 5	|T| = 7	|T| = 9
24	98.68%	98.08%	98.95%	98.59%	98.45%
48	98.92%	97.40%	96.83%	96.30%	96.89%
72	98.80%	98.34%	97.53%	98.01%	99.58%
96	99.54%	99.27%	99.40%	99.78%	100.55%
120	99.55%	99.46%	99.13%	101.49%	101.00%
GLA heuristic - maximum solution quality compared to Gurobi solution[1]					
|B|	|T| = 1	|T| = 3	|T| = 5	|T| = 7	|T| = 9
24	100.0%	99.98%	99.93%	99.87%	99.65%
48	100.0%	99.70%	99.32%	99.48%	99.67%
72	100.0%	99.80%	99.76%	99.82%	99.58%
96	100.0%	99.85%	99.60%	100.21%	100.55%
120	100.0%	99.67%	99.20%	102.45%	101.00%
GLA heuristic - minimum solution quality compared to Gurobi solution[1]					
|B|	|T| = 1	|T| = 3	|T| = 5	|T| = 7	|T| = 9
24	90.71%	91.04%	95.66%	95.27%	96.17%
48	96.52%	91.33%	94.19%	92.40%	94.24%
72	91.74%	96.21%	96.24%	96.04%	99.58%
96	98.38%	98.69%	98.99%	99.36%	100.55%
120	97.71%	99.05%	99.00%	99.57%	101.00%

[1] values below dashed line reflect the solution obtained when stopping the Gurobi solver after 12 hours

Table 4.6: Solution quality of GLA heuristic compared to Gurobi solution

the comparison between the values of the objective function based on the patient bed allocations created by both approaches. In particular, the GLA heuristic was able to achieve a solution quality of more than 95% for all comparable problem sizes. In addition, two effects can be observed from the data in Table 4.6. First, the average solution quality of the heuristic decreases slightly when increasing the planning horizon. This is to be expected since a longer planning horizon creates more favorable combinatorial combinations of patient bed allocations which are not straightforward and as such will likely not be detected by the heuristic approach. Second, the minimum solution quality and with it the average solution quality generally

increases the more beds are considered. This may be attributed to the higher probability of having comparable solutions in terms of the respective objective function value when increasing the number of beds. In other words, the heuristic will likely find a similarly adequate patient bed allocation even it moves away from the near to optimal solution. This can also be seen when comparing the solutions obtained by the Gurobi solver when stopped after 12 hours, i.e., the values below the dashed line. Here, the heuristic even outperforms the solution obtained by Gurobi's MIP solver for large problem instances while using only a fraction of the time. In summary, these analyses indicate that the use of the GLA heuristic developed may indeed deliver high solution quality results even for very large problem instances while at the same time providing ad-hoc online recommendations within seconds.

4.4.3 Case study

The modeling and solution approach suggested is applied at a large hospital in Germany. The first paragraph presents the data and parameters used followed by the second paragraph which presents the main results for our case study.

Environment, data used and methodology For our case study we investigated two departments covering 55 rooms with a total of 120 beds spread across 5 wards. This combination represents the pooled bed capacity for inpatients from the cardiology and gastroenterology departments.

We further obtained a detailed data set covering admission, discharge, as well as room transfer time stamps comprising the exact date and time on which each individual patient was actively booked in or out of a room and bed. In addition, the data set contains the department, age, gender, and care level of each individual patient and includes all data points recorded

for the cardiology and gastroenterology departments between January 2013 and September 2016. In this context, it is important to note that the available data only represents ex-post data, representing "what actually happened". However, in a real-life situation, it is often not clear before-hand how long a patient will stay as the anticipated discharge date is very likely to change throughout the stay of a patient. Thus, we tracked actual patient movements, as well as the actual predictions from physicians regarding the anticipated discharge date on site over the course of 4 weeks for all cardiology and gastroenterology patients on the associated wards. We then used these distributions and combined them with the ex-post data set at our disposal to prepare a series of event-based data points per patient which may be used to mimic all relevant information known to a potential bed planning system at a certain point in time. To this end, each data point comprises all patient-specific parameters and time stamps of anticipated arrivals or discharges as well as the exact time and date, on which these parameters and time stamps were last updated. Using the above-described approach, we apply nine data sets spanning from January 2016 to September 2016, with each set comprising all events, i.e., initial patients, admissions, discharges, and updates of LOS, occurring within a specified 28-day period. On average, every data set comprises 648 unique patients with around 2000 unique events taking place over the course of 28 days. This means that a deterministic problem instance was solved around 2000 times to simulate real-life application of our solution approach over time. The actual bed occupancy situation at the beginning of each time period is taken to initialize the calculation of each data set. All relevant patient parameters of current emergency and elective patient arrivals, future elective patient arrivals, as well as anticipated future emergency patient arrivals are taken into account to run the GLA heuristic. Due to the fact that almost all elective arrivals are known two weeks in advance, the time horizon taken into account for each run-through of the GLA heuristic was set to 14 days. In order to prevent overfitting, we used the available data from 2013 to 2015 to determine probability distributions for day-specific arrival rates, LOS, care level, department affiliation, and the age as well gender of emergency patients.

To test our modeling and solution approach, we compared the status quo, i.e., the actual PBA decisions taken in the case hospital, with the patient-bed allocations decisions that our GLA heuristic would have taken within the same time period. The GLA heuristic reruns every time a new event occurs, with the best known data currently available. After having undertaken all patient-bed allocations within the relevant data sets, the actual patient-bed allocations were analyzed ex-post facto based on the objective function of our decision model. We summarized the objective values as well as the relevant patient-specific, doctor-specific, and nurse-specific indicators and weighted them by the actual hours they were valid, respectively. In this context, it should be noted that this does not include the additional benefit OV_p attributed to patients coming from the overflow area, because it is merely an instrument to ensure that patients are not "left" in the overflow area. The "status quo", i.e., the ex-post evaluation of the patient-bed allocations that have actually taken place at the case hospital were used as a baseline. In addition, to assess the performance of our approach, we evaluated an "elective" scenario in which every future emergency patient and their characteristics as well as expected LOS are deterministically known beforehand, i.e., a scenario in which all patients are considered to be elective patients. However, the uncertainty of changes in LOS during the hospitalization are still existent in this scenario. Due to the remaining uncertainty, the result of the "elective" scenario is not necessarily better than the status quo. Nevertheless, it is expected that the "elective" scenario having significantly fewer uncertainty factors than the status quo achieves a higher solution quality.

Case study results Looking at the results of our case study in Table 4.7, the normalized values (normalized to the average of the status quo) of the objective function give a first indication of the performance of our approach. It can be noted that by using our GLA heuristic, i.e., the results in the columns termed "heuristic", it was possible to improve the patient-bed allocations for every available data set. In addition, the accumulated values of the objective function are fairly close for the "heuristic" and the "elective"

	accumulated OF values (normalized[1])			**utilization** (in percent)		
DS	status quo	heuristic	elective scenario	status quo	heuristic	elective scenario
1	98.3 %	103.0 %	103.3 %	76.5 %	82.0 %	81.9 %
2	100.9 %	105.3 %	105.7 %	77.8 %	82.9 %	82.9 %
3	98.9 %	103.4 %	103.6 %	76.3 %	82.5 %	82.5 %
4	106.6 %	111.5 %	112.0 %	81.9 %	88.0 %	88.0 %
5	102.7 %	106.1 %	106.5 %	78.4 %	82.8 %	83.2 %
6	100.4 %	103.3 %	103.3 %	77.6 %	82.0 %	81.9 %
7	102.0 %	106.1 %	106.2 %	78.7 %	83.4 %	82.8 %
8	94.0 %	97.1 %	97.4 %	72.1 %	76.9 %	76.9 %
9	96.2 %	99.6 %	98.8 %	73.5 %	77.9 %	77.5 %
∅	100.0 %	103.9 %	104.1 %	77.0 %	82.0 %	82.0 %

	overflow (in hours)			**age difference** (average in years)		
DS	status quo	heuristic	elective scenario	status quo	heuristic	elective scenario
1	2951	14	140	11.6	4.9	4.4
2	2960	158	209	11.1	5.0	4.3
3	3174	89	95	11.1	5.4	5.2
4	3407	147	430	12.1	5.5	5.4
5	2692	64	132	10.5	5.1	5.3
6	2625	142	117	11.3	5.6	5.3
7	2354	111	426	10.5	5.1	4.7
8	2408	115	115	10.1	5.1	4.8
9	2223	63	29	9.9	5.1	4.0
∅	2754	100	188	10.9	5.2	4.8

	same department (percentage of rooms)			**care level surplus** (average excess of threshold)		
DS	status quo	heuristic	elective scenario	status quo	heuristic	elective scenario
1	73.9 %	92.1 %	92.7 %	0.27	0.21	0.16
2	69.8 %	94.3 %	94.1 %	0.21	0.24	0.10
3	86.1 %	95.4 %	94.5 %	0.47	0.46	0.27
4	81.8 %	93.6 %	92.7 %	0.59	0.16	0.09
5	87.4 %	96.3 %	95.5 %	0.22	0.21	0.14
6	81.4 %	93.3 %	95.3 %	0.39	0.46	0.29
7	77.8 %	95.5 %	95.0 %	0.40	0.50	0.21
8	80.0 %	95.2 %	96.1 %	0.17	0.23	0.07
9	83.1 %	95.0 %	95.3 %	0.04	0.06	0.06
∅	80.1 %	94.5 %	94.6 %	0.31	0.28	0.15

[1] normalized to the average of the status quo values

Table 4.7: Case study analyses

scenario case. The application of our modeling approach significantly reduces the time patients have to spend in the overflow area and increases the average utilization. Utilization in this context is defined as the ratio of patients occupying a regular bed within their associated department-ward combination to the total number of beds within that department-ward combination. This is mainly due to three different effects with regard to the status quo. First, female and male patients are more efficiently combined to rooms such that situations in which several male-occupied rooms still have beds available while there is no room left for incoming female patients are prevented, and vice versa. Second, “standard patients” are less likely to block rooms and beds equipped with special infrastructure which they do not need. Third, medical isolation cases that may be combined, e.g., due to similar medical conditions, are more likely to be allocated to the same room instead of blocking multiple rooms. In comparison, the “elective” scenario actively generates slightly more overflow as all future emergency patients are already known which allows for trade-offs such that a slightly longer allocation of a patient to an overflow area may entail a better combination of patients in rooms and wards with regard to patient-specific, doctor-specific, and nurse-specific objectives. This is expected.

Furthermore, the results of our “heuristic” approach regarding average age difference, adherence to the same department, and care level all show significant improvements compared to the "status quo" scenario. In particular, it was possible to cut the average age difference in half while at the same time improving the percentage of rooms that only accommodate patients from a single department by 18 %. Finally, it was also possible to decrease the total amount of additional workload for nurses exceeding the respective predefined thresholds per ward by 10 %.

Traced run-times of the case study are demonstrated in Table 4.8. The average for each data set is built over all run-throughs during the period of 28 days. Any event that may change the planned patient-bed assignments,

	runtime [sec]		
	average	**minimum**	**maximum**
data set 1	1.026	0.213	1.557
data set 2	0.949	0.220	1.487
data set 3	0.955	0.259	1.249
data set 4	1.034	0.329	1.394
data set 5	0.904	0.289	1.378
data set 6	1.115	0.353	1.495
data set 7	1.017	0.348	1.317
data set 8	0.995	0.272	1.341
data set 9	0.923	0.195	1.410
∅	0.991	0.275	1.403

Table 4.8: Runtime analysis for one run-through

i.e. an emergency arrival or an update of the LOS, triggers a rerun of the GLA heuristic. This ensures a planning decision based on all information known to the system at that particular point in time. Each of the nine datasets averaged around 950 total run-throughs of the algorithm, i.e., complete patient-bed allocation updates. For each complete update the average runtime was less than one second and the maximum runtime does not exceed 1.6 seconds. It should be noted, that this runtime comprises the complete replanning effort, i.e., the assignment of all patients $p \in P$ to all available beds $b \in B$. In summary, the case study proves that the developed GLA heuristic is suitable and applicable as decision support system for the daily use in a large hospital.

4.4.4 General applicability

In addition, we investigated the general applicability of our proposed approach. Therefore, we drew on the data sets made publicly available by Demeester et al. (2010) and applied our approach for each data set. Although these data sets do not include previous occupancies, uncertainty, or care levels of patients, they do provide several large-sized problem instances which may be used for assessing computation times.

DS[1]	|P|	|B|	|T|	**utilization**	**age dif. [yr]**	**same dep.**	**run time [sec]**
1	693	286	14	60%	3	61%	2.5
2	778	465	14	60%	2.8	59%	4.7
3	757	395	14	57%	2.5	60%	3.8
4	782	471	14	54%	2.4	65%	4.7
5	631	325	14	49%	2.4	67%	2.5
6	726	313	14	64%	4.2	56%	3.5
7	770	472	14	34%	2.4	78%	1.5
8	895	441	21	44%	3.3	70%	4.3
9	1400	310	28	77%	11.2	40%	12.7
10	1575	308	56	48%	4.3	68%	17.6
11	2514	318	91	46%	3.6	71%	46.5
12	2750	310	84	55%	5.8	62%	55.8

[1] made publicly available by Demeester et al. (2010) for benchmarking purposes.

Table 4.9: General applicability analyses

Our results which can be seen in Table 4.9 show that the overall utilization in Demeester's data sets is so low that short-term allocations of patients to an overflow area are basically not required. Nonetheless, even for the largest problem instance, i.e., data set 12 comprising 2750 patients and a planning horizon of 84 days, we found a solution with less than a minute computation time. The achieved average age difference of the data sets varies around 3 years with one outlier of 11.2 years for data set 9. This is caused by the fact that data set 9 involves a pediatric and geriatric department as well as high utilization. Thereby, to avoid overflow situations, the model is forced to

combine patients with a large age gap. The percentage of patients adhering to the same department in one specific room per night ranges between 40% and 78%. This depends on the amount of specialities and the utilization.

In summary, the data sets provided by Demeester et al. (2010) significantly differ to what we encountered at our case hospital, in particular due to the unusually low utilization rates, as well as the lack of uncertainty in patient arrivals and updates of LOS. Thus, we additionally tested our model and solution approach with real-life data provided by the case hospital.

4.4.5 Sensitivity analyses

To better understand the trade-off effects that exist between the different objectives for patients, doctors, and nurses we created four additional scenarios in which we increased each of the four weighting factors α, β, γ, and δ by a factor of 10, respectively (see Table 4.10).

	base scenario	**scenario 1**	**scenario 2**	**scenario 3**	**scenario 4**
α	1	10	1	1	1
β	0.01	0.01	0.1	0.01	0.01
γ	0.2	0.2	0.2	2	0.2
δ	0.2	0.2	0.2	0.2	2

Table 4.10: Scenarios for sensitivity analyses

Each scenario is run with each of our nine real-life data sets. The results in Table 4.11 show the aggregated average values for each scenario and can be interpreted as follows. Throughout all four additionally created scenarios the utilization remains fairly constant around 82%. However, the individual results regarding overflow, age difference, department affiliation, and care level surplus show significant differences. This behavior is to be expected as the GLA heuristic will always try to fill up the available bed

capacities. Nonetheless, significant trade-offs can be seen when focusing on optimizing age differences, department adherence per room, and workload for nurses. For instance, it can be seen that focusing optimization on parameters with a higher variance such as the age of patients significantly increases overflow as patients are "held back" in the overflow area to achieve even better pairings with other patients in the future. On the other hand, optimizing parameters with a low variance, such as department adherence, does not have a measurable effect on overflow. Focusing on age difference or department adherence significantly reduces the performance of the respective other parameter, as both parameters are room-specific, meaning that a trade-off has to be found. By contrast, the overall workload for nurses is ward-specific. Thus, a strong focus on balancing this workload does not lead to significantly worse values regarding age difference and department adherence. Finally, a strong focus on weighting factor α leads to a decrease in all observed objectives. This is due to the fact that in such a constellation, the GLA heuristic always prefers incoming and future elective patients regardless of how good a current emergency patient might match with an elective patient when allocated to the same room, thus leading to a slightly higher overflow of emergency patients.

	utilization	**overflow**	**age difference**	**same department**	**care level surplus**
base scenario	82.0%	100	5.2	95%	0.28%
scenario 1	82.0%	84	6.3	90%	0.57%
scenario 2	81.2%	480	3.3	80%	0.25%
scenario 3	82.0%	67	11.5	98%	0.37%
scenario 4	82.0%	92	5.2	94%	0.13%

Table 4.11: Sensitivity analyses for patient-, doctor-, and nurse-specific objectives

4.5 Conclusion and further areas of research

Conclusion The present paper presents a decision model that can be applied for ad-hoc operational bed allocation in large hospital settings. Most of the previous literature on PBA focuses on the developed model of Demeester et al. (2010) and his published fictive example data sets, either by providing alternative and/or improved heuristic solution approaches for the problem defined and/or by adding certain aspects to the problem. In our joint project with a large German hospital covering all major disciplines, we identified a variety of additional aspects which to the best of our knowledge have not been dealt with in the literature currently available. Based on the real-life situation our decision support model incorporates three main stakeholders, namely patients, nursing staff, and doctors. The developed model integrates the planning of current emergency and elective patient arrivals, future elective patient arrivals, as well as anticipated future emergency patient arrivals. To the best of our knowledge, we are the first who take into account all relevant stakeholders, extended patient-patient room dependencies, overflow situations, and the anticipation of future emergency patients as well as the possibility of a frequent replanning, which accounts for the uncertainty being inherent in the system. The model and solution approach developed is designed to very quickly propose a meaningful bed allocation to the bed manager for every incoming patient at the time of their arrival, based on all the information known at that particular moment. We developed a greedy look-ahead (GLA) heuristic that is suitable and applicable for daily use as an efficient and quick support system. In the numerical results, we have shown that

i) the GLA heuristic greatly outperforms Gurobi's MIP solver in terms of computational time while delivering a solution quality of 96.8% and higher
ii) our GLA heuristic can also sufficiently solve large data sets from previous literature,

iii) on the basis of real hospital data the GLA heuristic improved the objectives of all stakeholders, e.g., the overflow was reduced by 96%,
iv) the objectives of the stakeholders are highly dependent on one another.

Finally, the modularity of our proposed approach regarding standard objectives and constraints of the typical stakeholders along with the ability to solve large problem instances renders our proposed approach applicable for large hospitals anywhere in the world which cater to most major disciplines and exhibit high emergency rates. As such it is not limited to the German setting.

Future areas of research Various opportunities exist for further research. Based on our decision model a survey on different sophisticated heuristics can be conducted, focusing in detail on the trade-off between runtime and solution quality. In addition, a more detailed investigation on the effects of uncertainty regarding emergency arrival ratios and LOS estimates can be undergone. This would include investigating different ways of modeling uncertainties for the multi-objective PBA problem. It is also imaginable to include further stakeholders such as, for example, bed transport services. The modeling and solution approach presented in this paper may be an appropriate starting point to address these open research questions.

5 Machine Learning and Pilot Method: Tackling Uncertainty in the Operational Patient-Bed Assignment Problem

Co-autors: Fabian Schäfer, Alexander Hübner, Dominik Grimm
submitted to *OR Spectrum* on 13 February 2020

Abstract This paper develops a multi-objective decision support model for solving the patient bed assignment problem. Assigning inpatients to hospital beds impacts patient satisfaction and the workload of nurses and doctors. The assignment is subject to unknown patient arrivals and lengths of stay, in particular for emergency patients. Hospitals therefore need to deal with uncertainty on actual bed requirements and potential shortage situations as bed capacities are limited. This paper contributes by improving the anticipation of emergency patients using machine learning approaches, incorporating weather data, time and dates, important local and regional events, as well as current and historical occupancy levels. Drawing on real-life data from a large case hospital, we were able to improve forecasting accuracy for emergency inpatient arrivals. We achieved an up to 17% better root mean square error when using machine learning methods compared to a baseline approach relying on averages for historical arrival rates. Second, we develop a new hyper-heuristic for solving real-life problem instances based on the pilot method and a specialized greedy look-ahead heuristic. When applying the hyper-heuristic in test sets we were able to increase the objective function by up to 3% in a single problem instance and up to 4% in a time series analysis compared to current approaches in literature. We achieved an improvement of up to 2.2% compared to a baseline approach from literature by combining the emergency patient admission forecasting and the hyper-heuristic on real-life situations.

5.1 Introduction

This paper deals with the patient bed assignment problem (PBA). This is the operational problem of allocating elective and emergency inpatients to specific rooms and beds within a hospital upon their arrival. The key challenge in PBA is the inherent uncertainty that governs most input parameters. The planning situation is unstable due to frequent changes, which may be caused by emergency patient arrivals, changes in treatment plans and a number of other factors. For example, large maximum care hospitals are a natural first point of contact for all emergency patients within their catchment area, which naturally leads to a high ratio of unknown emergency inpatient arrivals. Thus, when assigning inpatients to beds in such environments, it is very important to anticipate the number of imminent emergency patient arrivals as best as possible, as emergency and elective inpatients can occupy the same ward space. Several circumstances and external effects may drive the volume of emergency patients, e.g., seasons, weekdays, local events (e.g., county fairs, sports events). There may be different drivers for each discipline (e.g., snowy weather for trauma surgery, availability of family doctor for internal medicine). Real-life planning typically involves several hundred patients and beds, such that it is not uncommon to be faced with a completely changed set of input parameters due to several updates in the system during the planning horizon. Moreover, the PBA affects patient satisfaction (e.g., suitable room with adequate roommates), workload of nurses (e.g., a mix of work-intensive and easy-to-handle patients) and workload of doctors (e.g., own patients located in proximity). These may comprise some tradeoffs. For example, focusing only on patient satisfaction by putting optimal roommates together may be in conflict with the nurse workload. As such, the PBA is a multi-objective problem that considers the tradeoff between patient-, nurse-, and doctor-specific objectives while taking into account their respective constraints as well as infrastructural requirements.

The remainder of this paper is structured as follows: Section 5.2 defines the PBA problem, discusses related literature, and further elaborates on the specific contribution of this paper. Section 5.3 introduces the mathematical model and the hyper-heuristic framework developed. It is based on the "preferred iterative look-ahead technique" (pilot method) of Duin and Voß (1999) and Voß et al. (2005), which in part incorporates the greedy look-ahead heuristic described in Atkinson (1994) as a subheuristic. Section 5.4 provides several numerical examples based on actual hospital data and details a machine learning approach developed to better anticipate emergency inpatient arrivals. In addition, we combine these insights obtained from machine learning with a hyper-heuristic framework for solving the PBA efficiently for large problem instances. Finally, Section 5.5 presents a summary of the main results and outlines potential avenues for further research.

5.2 Problem description, related literature and contribution

5.2.1 General planning problem

Scope of the patient bed assignment problem It is important to distinguish between the patient admission and scheduling problem (PAS) and the patient bed assignment problem (PBA), as these expressions have been used with varying definitions in the literature. We consider the PAS as only dealing with the scheduling of elective patient admission dates (see e.g., Gartner and Kolisch (2014); Gartner and Padman (2019)), whereas the PBA tackles the problem of allocating a specific room and bed to a specific inpatient (see e.g., Demeester et al. (2010); Ceschia and Schaerf (2011); Schäfer et al. (2019)). For the PAS it is not necessary to know which bed exactly will be held available for a certain inpatient as long as it is guaranteed to a certain extent that a bed will be available (see e.g.,

Ceschia and Schaerf (2016)). The PBA is the downstream decision with regard to the PAS.

Figure 5.1 presents an example of the PBA. Two female emergency patients who have just arrived are planned to stay in beds 3 and 4. While bed 1 is theoretically available before bed 3, it is already "reserved" for a male elective patient scheduled to arrive on Friday and stay for several days. Consequently, the female patient planned to occupy bed 3 will have to wait in an overflow area (e.g., hallways, emergency or treatment rooms) until Saturday when bed 3 becomes available for her. For this example, it is considered more important that the elective patient arriving on Friday does not have to wait in an overflow area. Hence, it is crucial to determine at which time a specific physical room and bed is to be assigned to a inpatient and whether or not it should be possible to reserve such a bed. In essence, there is always a tradeoff between different PBAs, which at times leads to situations where it may be beneficial to the overall utility to deviate from a first-come-first-served rule.

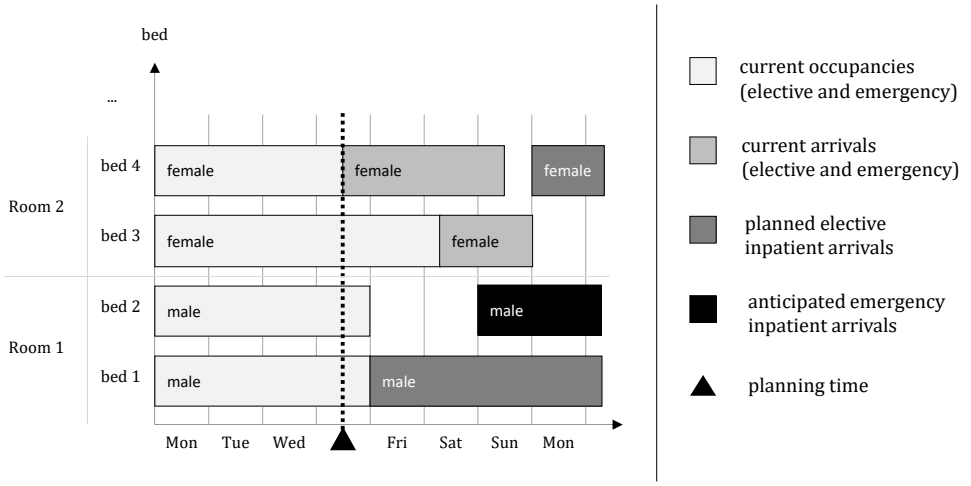


Figure 5.1: Illustration of the patient bed assignment problem

Objectives of patient bed assignment problem In general, patients want to have a room within a designated ward space that caters to their

medical needs while avoiding unnecessary room transfers or having to wait in an overflow area. It is further desirable to combine similar patients, e.g., patients of similar age or with similar illnesses in terms of their specific medical condition and the severity thereof. In addition, elective patients typically do not accept that a room and bed within their respective department is not "reserved" for them upon their arrival, while emergency patients are more willing to accept having to temporarily stay in dedicated overflow areas. If staying in an overflow area does become necessary, patients wish to be transferred to a "regular room" as soon as possible. To facilitate doing rounds and patient visits, walking distances for doctors should be minimized. This can be achieved by grouping similar patients, i.e., patients associated with a specific department, into rooms. Compared to doctors, nurses can typically tend to a broader range of patients. However, they are typically dedicated to a specific ward, working in well-coordinated teams, and therefore cannot easily be transferred to other wards. Thus, balancing workload between wards is a key objective for nurses when assigning patients to beds (Schäfer et al., 2019).

Constraints of the patient bed assignment problem For the PBA, the following conditions have to be taken into account. First, non-ICU (intensive care unit) female and male inpatients are not allowed to be allocated to the same room. Second, certain medical conditions require patients to be in rooms, that are equipped with the necessary infrastructure, e.g., telemetry for selected cardiology patients. Third, it may be the case that a patient or several patients need to be isolated from other patients during their stay for medical reasons. Finally, there are usually no non-medically induced room transfers, meaning that assignments of patients who already physically occupy rooms associated with their designated department are treated as unchangeable. This is due to the fact that every physical room transfer entails significant additional work for hospital personnel (e.g., cleaning and sanitizing rooms, moving beds, reorganizing tasks) as well as unnecessary discomfort for the patient. In this context,

the only exceptions are transfers due to medical reasons (e.g., transfers to and from the ICU).

5.2.2 Complexity of the patient bed assignment problem

In order to guarantee patient satisfaction and trouble-free process flow (i.e., avoid waiting times until inpatient admission as well as blocking emergency departments), bed mangers need real-time decision support. Furthermore, real-life planning situations are affected by many sudden changes (e.g., update of length of stay (LOS), no-shows and emergency patient admissions). Large hospitals in particular therefore require highly flexible planning systems, that are able to adapt to unexpected changes in real time. PBA complexity thus results from (1) being unable to precisely estimate the number of beds required and (2) the size of the problem of jointly planning hundreds of beds.

(1) Arrival and length of stay of patients Usually elective and emergency inpatients share the same ward space and bed capacities. This requires jointly planning the PBA for both types. Emergency inpatient arrivals are not known in advance and are stochastic, so they can only be estimated. Appropriately predicting which kind of emergency patients and how many are likely to arrive on a given day is a fundamental input to the PBA, particularly for large maximum care hospitals where up to 80% may be emergency patients. Simply predicting emergency patients based on historical averages will fall short, as – in addition to an inherent randomness – it seems highly probable that the actual number of emergency arrivals is dependent on a plethora of factors internal and external to the hospital, and cannot be explained solely by the time and date. For example, trauma surgery departments may experience an increase in emergency inpatients at the beginning of the cold season due to sidewalks that have

frozen over, leading to more elderly people falling down and suffering a fracture. Furthermore, the LOS of a patient is always an informed estimate. Unforeseeable events such as sudden complications during surgery or treatment, faster recoveries, or patients who self-discharge against medical advice can potentially lead to a change in the LOS. Some disciplines exhibit high emergency arrival rates or are subject to more LOS updates. Finally, elective patients can also fail to show up for their planned inpatient stay. All this together leads to high volatility regarding future occupancy levels. In combination with the economic need for tight capacity and high occupancy levels, the volatility in patient volume inevitably leads to occasional overflow situations. In such cases, inpatients need to be temporarily assigned to overflow areas. Hallways, emergency or treatment rooms, or other wards outside a dedicated ward space may serve as buffers in such cases. Staying in such intermediate areas, however, is unpleasant for patients and will always entail additional work for nursing staff and doctors alike.

(2) Size of the problem To obtain better capacity utilization, departments of large hospitals now share ward space (see e.g., Van Essen et al. (2015) and Hübner et al. (2018)), which calls for jointly planning hundreds of beds and efficient decision support. Whenever an elective or emergency patient is admitted to or discharged from wards, LOS are changed or no-shows of elective patients occur, or patients are reassigned from the overflow, the PBA needs to be updated. As such, the underlying planning problem has to be solved many times per day. To illustrate, one can for instance assume a scenario comprising a pooled capacity of 1,000 beds exhibiting an average utilization of 90%, with patients who stay three days on average. This would lead to an average of 300 inpatient arrivals and 300 inpatient dismissals per day, respectively. Further assuming an emergency ratio of 50% would mean that at least 150 of said arrivals are subject to fluctuations to the bed planning system beforehand. In addition, one can for example assume that 50% of the remaining elective PBAs, i.e., 75 arrivals, are somehow affected by a sudden change in LOS of any of the current occupants. In total, this would lead to an average of 225 additional events

during the day, for which all future PBAs have to be recalibrated. Further changes in LOS updates for patients already occupying a room, no-shows of elective patients and overflow situations will increase the number of events. In overflow situations unexpected inpatient dismissals could then directly affect potential PBAs of any patients currently waiting within an overflow area.

5.2.3 Related literature

The problem at hand is related to decision models for the PBA and relies on estimating emergency patients. We structure the literature review in these two areas, and derive the associated open research areas in each section.

5.2.3.1 Decision models and related literature for patient bed assignment

The PBA has gained more and more attention mainly within the past decade. Key challenges dealt with in most contributions to this area of research can be seen in the computational complexity of typical problem sizes and the resulting need for heuristic solution approaches, as well as the underlying uncertainty and volatility of most parameters involved. Table 5.1 gives an overview of most of the recent contributions and highlights a set of key aspects related to the challenges mentioned above. With regard to the modeling approach followed, “static setting” refers to a hypothetical scenario in which every future arrival is known and no prior occupancies are considered, whereas in the “dynamic setting” prior occupancies are considered, while future arrivals are only known within a defined planning horizon. Column “emergency patients” indicates whether or not the potential arrival of inpatients is considered, which cannot be known in advance. Furthermore, under “overflow possible” we indicate whether a specific modeling approach is designed to deliver solutions for situations when not enough beds are

available for all arriving inpatients. “Uncertainty considered” refers to the modeling of volatility in patient parameters, e.g., future changes in LOS. The column “stakeholders” indicates for which group, i.e., patients, nurses, and doctors, are included within the model. The column “emergency forecast” indicates whether emergency inpatient arrivals were analyzed beyond the effects of using simple historical occupancy distributions. Furthermore, the column “time series” indicates whether the continuous application of a PBA algorithm over the course of several days or weeks was analyzed. In essence, this means analyzing the actual accumulated partial benefits or costs incurred by each patient stay in retrospect. Finally, the column “data sets used” indicates whether simulated data or real-life hospital data was used.

Contribution	Basic solution approach	Modeling approach					Analyses		
		Problem setting[1]	Emergency patients	Overflow possible[2]	Uncertainty[3]	Stakeholders[4]	Emergency forecast[5]	Time series[6]	Data sets used[7]
Demeester et al. (2010)	Tabu search	S	–	–	–	P	–	–	S
Bilgin et al. (2012)	Hyper-heuristic	S	–	–	–	P	–	–	S
Kifah and Abdullah (2015)	Great deluge	S	–	–	–	P	–	–	S
Turhan and Bilgen (2017)	Fix-and-optimize heuristic	S	–	–	–	P	–	–	S
Guido et al. (2018)	Matheuristic	S	–	–	–	P	–	–	S
Bastos et al. (2019)	Exact approach	S	–	–	–	P	–	–	S
Dorgham et al. (2019)	Genetic algorithm	S	–	–	–	P	–	–	S
Taramasco et al. (2019)	Metaheuristics	S	–	–	–	P	–	–	R
Ceschia and Schaerf (2011)	LNS	S,D	–	–	✓	P	–	–	S
Ceschia and Schaerf (2012)	LNS	S,D	✓	–	✓	P	–	–	S
Ceschia and Schaerf (2016)	LNS	S,D	✓	–	✓	P	–	–	S
Lusby et al. (2016)	Adaptive LNS	D	✓	–	✓	P	–	–	S
Vancroonenburg et al. (2016)	Specialized heuristic	S,D	✓	–	✓	P	–	–	S
Schäfer et al. (2019)	Greedy look-ahead heuristic	S,D	✓	✓	✓	P,N,D	–	✓	S,R
This Paper	Hyper-heuristic based on Pilot; ML for emergency forecast	S,D	✓	✓	✓	P,N,D	✓	✓	R

1 S = static version of the PBA (every future arrival known, no prior occupancies)
D = dynamic version of the PBA (prior occupancy considered, arrivals known within planning horizon)
2 Situations in which not enough beds are available and in which patients have to spend time in an overflow buffer until a bed becomes available
3 uncertainty regarding patient LOS and future admission dates
4 objectives and constraints considered for patients (P), nurses (N), and doctors (Doc)
5 emergency inpatient arrival forecast (analysis of effects beyond using simple historical averages)
6 time series version of the PBA (analysis of effect that continuous application of a PBA-algorithm has over the course of several weeks)
7 S = simulated problem instances, R = real hospital data

Table 5.1: Overview of decision models related to patient bed assignment

Static models for patient bed assignments The PBA was first introduced by Demeester et al. (2010). They consider a situation in which a hospital is initially empty and all future patient arrivals within a given time horizon are deterministically known as well as their respective parameters, e.g., actual LOS, gender, department adherence, individual infrastructural needs. The model is formulated as a static model where the assignments are only made once to populate the hospital. This static version is only a single problem instance in which all future arrivals are known, without considering any prior occupancies. In their model, patients are assigned to rooms such that an overall objective function based on violating patient-specific requirements is minimized. The model acknowledges gender-specific room assignment, assignment of patients to departments suited to their age, availability of relevant infrastructure, adherence to medical isolation and patient-specific room type preferences (e.g., single or double room). Patients are assigned to available rooms of a certain type while taking known admission and discharge dates of each patient into account. Capacity is assumed to be sufficient to accommodate all inpatients. As such, it does not allow for overflow situations, i.e., problem instances in which not enough beds are available for all inpatients cannot be solved. Furthermore, they do not consider nurse- and doctor-specific objectives and do not distinguish between emergency and elective patients. They apply a token-ring tabu search.

Several authors have since built on the model developed by Demeester et al. (2010) by providing alternative or improved solution approaches and/or by introducing new aspects to the PBA. Bilgin et al. (2012) use the model provided by Demeester et al. (2010) and solve the static version of the PBA by applying a hyper-heuristic approach using simulated annealing and a tabu search. Kifah and Abdullah (2015) and Bastos et al. (2019) also provide new solution approaches to the static version of the PBA model as proposed by Demeester et al. (2010). In particular, Kifah and Abdullah (2015) propose a variant of a generic algorithm, i.e., an adaptive non-linear great deluge heuristic, whereas Bastos et al. (2019) propose an MIP formulation and use sparcity conditions to find optimal solutions

for some problem instances of Demeester et al. (2010). To decrease the computational complexity, Ceschia and Schaerf (2011) have proposed a reformulated version of the mathematical model originally proposed by Demeester et al. (2010). Specifically, Ceschia and Schaerf (2011) reformulate the model such that patients are only assigned to rooms rather than beds, as they consider the beds in each room to be identical. Based on this reformulated version, Turhan and Bilgen (2017), Guido et al. (2018), and Dorgham et al. (2019) have presented solution approaches to the PBA. For instance, Turhan and Bilgen (2017) also focus on improving the static version of the PBA problem and investigate the effects of using a fix-and-optimize heuristic. In addition, Guido et al. (2018) further investigate the impact of switching hard and soft constraints in the PBA and develop a matheuristic that focuses on providing tighter bounds on the search space. More recently, Dorgham et al. (2019) have proposed a further variant of a genetic algorithm combined with a hybrid simulated annealing approach. Taramasco et al. (2019) on the other hand have taken a slightly different modeling approach to the PBA. Specifically, they investigate a network of hospitals and divide the PBA into two stages. In a first stage patients are assigned to beds within a specific hospital, while in a subsequent second stage patients who cannot be assigned an adequate bed are redistributed among the other hospitals within the network. In addition, Taramasco et al. (2019) are one of the few who have investigated the static version of the PBA using real-life hospital data. To solve their model for large problem instances, they propose a metaheuristic, which is a composition of different specialized and evolutionary heuristics and approximate methods.

Investigating the static version of the PBA provides a valuable controlled test environment, which has been used in the literature to compare different modeling and solution approaches. However, for real-life applicability, a proposed modeling and solution approach for the PBA needs to be able to handle dynamic online planning situations, i.e., problem instances in which some beds are already pre-occupied and in which not all future arrivals are fully known to the system. In addition, for large hospitals using pooled ward capacities and experiencing high ratios of emergency arrivals, it is

especially important to incorporate potential emergency inpatients into the bed assignment planning. First and foremost this requires having good emergency arrival forecasts. Furthermore, when pairing high occupancy rates with high emergency arrival rates, overflow situations are likely to arrive that have to be handled by the PBA system.

Dynamic models for patient bed assignments Ceschia and Schaerf (2011) are the first to provide an approach for adapting the PBA model and solution approach to the dynamic case. To this end, they include the notion of an individual "registration date" per patient, i.e., the date the arrival of the patient becomes known to the system. The number of days an arrival is known in advance can vary for elective patients and can be considered to equal zero for emergency arrivals. However, emergency patients are not treated any differently to elective patients once they are known to the system. In addition, Ceschia and Schaerf (2011) consider pre-occupancies, i.e., patients who are already in the hospital at the planning date, whereby each PBA that has happened before said date is considered fixed. To test their approach they draw on simulated data by Demeester et al. (2010) and adapt the information in a reasonable but arbitrary way. Furthermore, Ceschia and Schaerf (2011) provide an approach to investigating the uncertainty regarding the discharge date of a patient that is inherent in the dynamic problem setting. To assess the impact of different LOS they solve the PBA several times using different values for the discharge dates of all patients in the system. Specifically, they start by assuming that each patient leaves after one day and add a day to the LOS of each patient, respectively, until the actual discharge date is reached. In their subsequent work (Ceschia and Schaerf (2012) and Ceschia and Schaerf (2016)) they further include uncertainty by factoring in flexible horizons and patient delays while also adding operating room constraints. Based on the work of Ceschia and Schaerf (2012), Lusby et al. (2016) further provide an alternative solution method to the PBA under uncertainty. Specifically, they develop an adaptive search procedure. Vancroonenburg et al. (2016) tackle the dynamic PBA setting by providing a first model that is designed

to only assign those patients to a new room who have just arrived and physically require a bed. They use a graph-based approach in which they use maximal cliques to respect room capacity constraints. In addition, they suggest a second model in which they also assign patients to beds who are registered in the system but have not yet arrived. This approach uses "dummy rooms" that are only "open" to patients who have not yet arrived in order to ensure feasibility of the model in undercapacity situations. Schäfer et al. (2019) have developed a comprehensive model and a specialized solution approach for solving the PBA. Their model distinguishes between emergency and elective patients and incorporates their respective needs and constraints as well as those of doctors, nurses, and management. In addition, it is designed to handle ad-hoc overflow situations, should they arise. Finally, it incorporates and evaluates patient-patient dependencies with regard to rooms and wards.

For real life situations in large hospitals, it is important to have a decision support system that is proven to work in a dynamic online scheduling scenario. At a minimum, this requires a solution approach that can deal with ad-hoc overflow situations and emergency inpatient arrivals. In addition, the underlying volatility of patient LOS and emergency arrival rates typically requires several adaptations of future PBAs during any given day. The performance of any such support system can thus only be measured by retrospectively evaluating actual occupancy. To the best of our knowledge, Schäfer et al. (2019) are the only ones to have analyzed the performance of their modeling and solution approach over a time series. However, there is still a need for better parameter forecasting for the dynamic problem setting and testing.

Further literature related to patient bed assignment To complete the picture, we additionally review related literature to highlight further aspects, that are considered relevant to the PBA context. For instance, bed capacity related issues are addressed by the following authors. Van Essen et al. (2015) and Hübner et al. (2018) develop approaches to combine

departments and wards to pool bed capacities. Vanberkel et al. (2012) use a queuing model to investigate the tradeoffs between centralizing hospital resources and decentralizing. Holm et al. (2013) use a discrete event simulation model to analyze patient flows and optimize the assignment of bed capacities between wards. Bekker et al. (2016) investigate the issue of partially flexible ward capacity and how much should be attributed to a general overflow area. Another example for handling overflow situations can be found in Herring and Herrmann (2012) who investigate the effects of deferring surgical patients while blocking surgical capacity for higher priority cases. Cotta (2011) investigate the effects of patient prioritization in a mass casualty scenario. With regard to patient admission, for instance, Gartner and Padman (2019) build on and extend Gartner and Kolisch (2014)'s approach to solving the PAS. They focus on the assignment of hospital resources and provide a mathematical program, that, among other things, includes flexible patient assignments to medical departments to account for multi-morbid patient clientele, as well as overtime availability of medical and nursing staff. Luscombe and Kozan (2016) provide a dynamic scheduling framework that relates to parallel machine and flexible job shop problems to provide a decision support model for patient assignment in emergency departments.

Open research with regard to modeling and solving PBAs As to the operational assignment of beds, the actual problem at hand is a dynamic online planning situation in which the PBA needs to be solved several times per day. That means at each point in time that an inpatient gets admitted or discharged or when any other change in the system merits moving patients from an overflow area to a regular bed. Alternative heuristics are required to address the dynamic problem. As pointed out above, a key performance indicator for any such heuristic is the retrospective assessment of actual occupancies over time. To the best of our knowledge, Schäfer et al. (2019) are the only ones to have provided such a time-series analysis using a deterministic greedy look-ahead heuristic. However, there is still a need to

investigate more sophisticated heuristic approaches using different parameter settings, especially when using non-deterministic solution approaches.

5.2.3.2 Literature related to estimating emergency patients

One of the key drivers of uncertainty regarding bed management in large hospitals is the large ratio of emergency inpatient arrivals, which for certain medical specialties such as cardiology can surpass 80%. Carvalho-Silva et al. (2018) as well as Afilal et al. (2016) concern themselves with the problem of forecasting emergency arrivals at a hospital. Both use hospital data and use an autoregressive moving average approach. Schiele et al. (2019) provide a model to anticipate resulting bed occupancy levels based on a given master surgery schedule. They consider different patient types and paths and make use of a neural network based approach to improve their prediction quality. In addition, several authors have dealt with forecasting emergency arrivals in general, e.g., outpatient arrivals, day clinic walk-ins, or emergency calls, as can be seen in a systematic review written by Wargon et al. (2009). More recently, Gul and Celik (2018) have reviewed and analyzed contributions on applications of statistical forecasting in emergency departments.

Open research with regard to estimating emergency patients for PBA As pointed out above, anticipating emergency arrivals as accurately as possible is key for the PBA. Our literature review shows that advanced methods to better anticipate emergency inpatient arrivals, e.g., using state-of-the-art machine learning methods, are rare in general and not available for the specific problem of assigning inpatients to beds. To this end, a broader investigation with combined effects such as detailed weather data, holidays, seasons or significant local events is required. This will allow the prediction of emergency arrivals more accurately compared to solely drawing on historical averages and distributions of patient arrivals. Such an approach is promising as it relies on publicly available data and as such is possible to be incorporated in existing planning systems. To the best of our

knowledge, such an integrated approach to forecasting emergency inpatient arrivals for the PBA has not yet been proposed in the literature.

5.3 Modeling and solution approach

5.3.1 Model complexity, general idea of the solution approach and model overview

Complexity and general idea The underlying problem of the PBA could be represented as a stochastic dynamic program. The dynamic setting of the problem arises from multiple events such as arrivals, discharges and no-shows of patients as well as changes in LOS. Here, each event represents a stage and the total number of inpatients constitutes the state space in each stage. To illustrate, when assuming the case of a large hospital with about 800 beds occupied on average, a planning horizon of 28 days and an average of over 500 events per day, this would result in more than 14,000 stages and a total state space of more than 11 million entries. The stochastic volatility arises from the fact that the total number and type of inpatients cannot be predetermined and are further subject to uncontrollable external influences (such as weather, patient recovery, treatment complications, etc.). In light of this, it becomes obvious that such a dynamic problem setting cannot be solved optimally, meaning that a heuristic approach is required if one wants to provide efficient and effective decision support in real-life settings. We approximate the dynamic problem as Schäfer et al. (2019) by solving a static model that is updated at each possible event. Ceschia and Schaerf (2011) propose a similar approach to test the performance of their static model in a dynamic setting. When solving the model, it allocates beds for patients (new inpatients and patients from overflow buffer), assigns patients to overflow, and reserves beds for patients (currently in overflow and future patient arrivals). As such, we subsequently solve single stages while considering future arrivals and discharges that are both already known and

estimated. The model takes all the relevant information currently available into account for each of these individual stages.

Model overview The decision model is based on Schäfer et al. (2019). A multi-objective utility maximization problem quantifies patient-specific, doctor-specific, and nurse-specific objectives, while simultaneously considering medical, gender, and capacity constraints. The model builds in the possibility of using a buffer for situations where the number of beds is insufficient or beds may be blocked for patients arriving later. Table 5.2 summarizes the notation.

Sets	
B	Set of beds, $B = \{1, 2, ..., b, ..., \|B\|\}$
D	Set of medical departments, $D = \{1, 2, ..., d, ..., \|D\|\}$
P	Set of inpatients, $P = \{1, 2, ..., p, ..., \|P\|\}$
R	Set of rooms, $R = \{1, 2, ..., r, ..., \|R\|\}$
T	Set of days within the planning horizon, $T = \{1, 2, ..., t, ..., \|T\|\}$
W	Set of wards, $W = \{1, 2, ..., w, ..., \|W\|\}$
Parameters	
$\alpha, \beta, \gamma, \delta$	Weights for basic and extended patient-, doctor- and nurse-related utilities, respectively
Ξ_p	Weight for patients o (e.g., elective vs. emergency patient)
a_p	Age of patient p
$A_{rt}^{\max}$ $\left(A_{rt}^{\min}\right)$	Maximum (minimum) age of all patients already occupying room r on day t
C_{wt}	Spare care capacity for caring for further patients on ward w on day t
c_p	Care level required to accommodate patient p
D_{rt}	1 if all prior occupants of room r on day t belong to same medical department; 0 otherwise
d_p	Medical department of patient p with $d_p \in D$
E_{bt}	1 if bed b is located in a room that is initially empty on day t; 0 otherwise
F_{rt}	1 if room r is initially empty on day t; 0 otherwise
g_p	-1 if patient p is male; 1 if patient p is female
i_p	$i_p = -1$ if patient p requires medical isolation; 1 otherwise
K_{br}	1 if bed b is in room r; 0 otherwise
L_{bw}	1 if bed b is in ward w; 0 otherwise
OF_p	Utility parameter of patient p depending on the time patient p has already spent in overflow

Continued on next page

Table 5.2 – *Continued from previous page*

Q_t	Time-dependent relevance value that arrivals/discharges will take place as anticipated/planned on day t
s_{bpt}	1 if bed b is available for patient p on day t; 0 otherwise
Decision variable	
x_{bp}	1 if patient p is assigned to bed b; 0 otherwise
Auxiliary variables	
$a_{rt}^{\max}$ ($a_{rt}^{\min}$)	Maximum (minimum) age of all patients p assigned to room r on day t
o_{wt}^{+}	Amount the total care capacity on ward w on day t is exceeded
y_{rt} (z_{rt})	1 if all patients assigned to an empty (partially occupied) room r on day t are from the same medical department; 0 otherwise

Table 5.2: Notation

The objective function of Equation (5.1) maximizes the total utility U and consists of four terms that represent basic patient-specific objectives, extended patient-specific objectives, doctor-specific objectives and finally nurse-specific objectives. The four partial utilities are weighted by the factors α, β, γ, and δ. All four utility values depend on the binary assignment variable x_{bp} that represents whether a patient $p, p \in P$ is allocated to bed $b, b \in B$. The model is formulated as follows:

$$\begin{aligned}\text{maximize } U = \alpha \sum_{b\in B}\sum_{p\in P}(\mathrm{OF}_p + \Xi_p \sum_{t\in T} s_{bpt}Q_t)x_{bp} - \beta \sum_{r\in R}\sum_{t\in T}(a_{rt}^{\max} - a_{rt}^{\min}) \\ +\gamma \left[\sum_{r\in R}\sum_{t\in T} f_{rt}y_{rt} + \sum_{r\in R}\sum_{t\in T}(1-f_{rt})z_{rt}\right] - \delta \left(\sum_{w\in W}\sum_{t\in T} o_{wt}^{+}\right)\end{aligned} \tag{5.1}$$

subject to

$$\sum_{b \in B} x_{bp} \leq 1 \qquad \forall p \in P \tag{5.2}$$

$$\sum_{p \in P} s_{bpt} x_{bp} \leq 1 \qquad \forall b \in B; t \in T \tag{5.3}$$

$$\sum_{t \in T} s_{bpt} \geq x_{bp} \qquad \forall b \in B; p \in P \tag{5.4}$$

$$K_{br} E_{bt} g_p s_{bpt} x_{bp} - K_{lr} E_{lt} g_h s_{lht} x_{lh} \geq -1 \qquad \forall b, l \in B; p, h \in P; r \in R; t \in T \tag{5.5}$$

$$K_{br} E_{bt} i_p s_{bpt} x_{bp} - K_{lr} E_{lt} i_h s_{lht} x_{lh} \geq -1 \qquad \forall b, l \in B; p, h \in P; r \in R; t \in T \tag{5.6}$$

$$a_{rt}^{\max} \geq A_{rt}^{\max} \qquad \forall r \in R; t \in T \tag{5.7}$$

$$a_{rt}^{\max} \geq K_{br} a_p s_{bpt} x_{bp} \qquad \forall b \in B; p \in P; r \in R; t \in T \tag{5.8}$$

$$a_{rt}^{\min} \leq A_{rt}^{\min} \qquad \forall r \in R; t \in T \tag{5.9}$$

$$a_{rt}^{\min} \leq \sum_{b \in B} \sum_{p \in P} A_{rt}^{\min} K_{br} s_{bpt} x_{bp} \qquad \forall r \in R; t \in T \tag{5.10}$$

$$a_{rt}^{\min} \leq K_{br} a_p s_{bpt} x_{bp} + A_{rt}^{\min}(1 - x_{bp}) \qquad \forall b \in B; p \in P; r \in R; t \in T \tag{5.11}$$

$$K_{br} d_p s_{bpt} x_{bp} - K_{lr} d_h s_{lht} x_{lh} \geq -\mathrm{M}(1 - y_{rt}) \qquad \forall b, l \in B; p, h \in P; r \in R; t \in T \tag{5.12}$$

$$\sum_{p \in P} \sum_{b \in B} K_{br} s_{bpt} x_{bp} \geq y_{rt} \qquad \forall r \in R; t \in T \tag{5.13}$$

$$K_{br} d_p s_{bpt} x_{bp} - D_{rt} \leq \mathrm{M}(1 - z_{rt}) \qquad \forall b \in B; p \in P; r \in R; t \in T \tag{5.14}$$

$$D_{rt} - K_{br} d_p s_{bpt} x_{bp} \leq \mathrm{M}(1 - z_{rt}) \qquad \forall b \in B; p \in P; r \in R; t \in T \tag{5.15}$$

$$\sum_{p \in P} \sum_{b \in B} K_{br} s_{bpt} x_{bp} \geq z_{rt} \qquad \forall r \in R; t \in T \tag{5.16}$$

$$\sum_{b \in B} \sum_{p \in P} L_{bw} c_p s_{bpt} x_{bp} \leq C_{wt} + o^{+}_{wt} \qquad \forall t \in T; w \in W \tag{5.17}$$

$$o^{+}_{wt} \geq 0 \qquad \forall t \in T; w \in W \tag{5.18}$$

$$x_{bp}, y_{rt}, z_{rt} \in \{0, 1\} \qquad \forall b \in B; p \in P; r \in R; t \in T \tag{5.19}$$

The first term of the objective function in Equation (5.1) summarizes the basic patient-specific utility of assigning patient $p, p \in P$ to bed $b, b \in B$. Every assignment of a patient p to a bed b, i.e., $x_{bp} = 1$ generates a utility that accounts for the days that patient p is presumed to spend in bed b within the planning horizon T. The utility depends on the time the patient p already spent in the overflow (OF_p) in the past, a patient type-specific factor (Ξ_p), bed availability (s_{bpt}), and a relevance value (Q_t). The incorporation of an overflow value in the first part of the utility function allows patients already waiting in the overflow area to be assigned a higher preference than similar patients who have just arrived in the hospital. The second part of the utility function rewards the actual time that a patient p spends in bed $b \in B$. The parameter Ξ_p is a factor that makes it possible to

distinguish between patient types, i.e., elective patients, emergency patients, or patients with special requirements. This factor may, for example, be used to ensure that elective patients are more likely to be assigned to a bed within their target ward upon arrival than emergency patients. In addition, patients returning from the ICU could be attributed an even higher value such that they will not be moved to an overflow area. The parameter s_{bpt} is set to 1 in the event that bed b is available for patient p on day t, and 0 otherwise. As non-medical room transfers are not allowed, s_{bpt} is determined at each event and is used to reflect not only bed availability but also bed compatibility by incorporating gender constraints (with respect to current occupants), infrastructural constraints, as well as medical isolation constraints (with respect to current occupants) for each possible patient bed combination. Figure 5.2 shows an example illustrating how parameter s_{bpt} is determined. The upper part represents the current occupancy and the lower part the determination of s_{bpt}. The parameter $s_{bpt} = 1$ if the respective bed is available for this patient on this day, otherwise there is no entry, meaning that $s_{bpt} = 0$. The example considers four rooms, each with two beds. A new female patient arrives on day 3 and is scheduled to be discharged on day 8. There are multiple options for allocating her to a bed. Male patients occupy room 1 with beds 1 and 2. Currently, the earliest availability of bed 1 and 2 for a female patient is day 6 after patient in bed 1 leaves. Therefore there are no entries in s_{bpt} for days 1 to 5. As she is scheduled to leave on day 8, day 8 to the end of the planning horizon has also no entry. Hence, assigning her to room 1 would result in spending at least two days in the overflow area. Bed 3 is available from day 4 and bed 4 is directly available. Beds 5 and 6 are not allowed to be used by this inpatient as this room is not equipped with essential medical infrastructure specifically required for this patient. Finally, female patients currently occupy both beds 7 and 8. They require medical isolation from non-quarantined patients for the duration of their stay, hence forcing the new patient to spend one day in the overflow area before moving into either of these beds, should she be allocated to one of them.

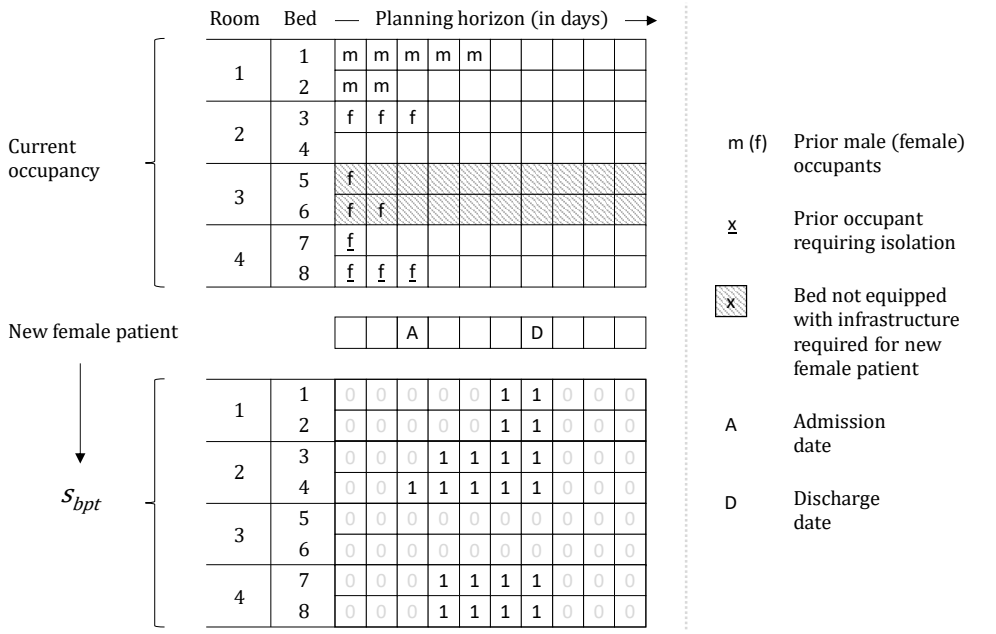


Figure 5.2: Example for determining parameter s_{bpt} for a new female patient arrival

Finally, Q_t is a parameter that reflects the time-dependent relevance of a bed assignment for patients on day t as anticipated/planned, where Q_t decreases with increasing t. It gives a higher value to patients who arrive earlier than those who come later in the planning horizon. Due to uncertainties it is quite reasonable that a patient who is planned to arrive far in the future will be reassigned to another bed during later planning periods, which may even lead to a higher overall utility value for that patient. Equations (5.2) prevent double booking, i.e., a patient can only be allocated to a maximum of one bed. Equations (5.3) prevent overbooking, i.e., no two patients can be allocated to the same bed on the same day. Equations (5.4) ensure that a patient p can only be assigned to a bed b if bed b is available for this specific patient on at least one day during the stay, i.e., $s_{bpt} = 1$ for at least one $t \in T$. In addition, Equation (5.5) ensures that there are no mixed male and female rooms on any given day t. Using a similar approach, Equation (5.6) ensures that medical isolation requirements are respected. Specifically, patients that need to be isolated due to infectious diseases, for example,

may only be put into empty rooms or into rooms with patients that suffer from the same condition.

The second term of Equation (5.1) represents the extended patient-specific part. It evaluates the compatibility between different patients occupying one room. The goal is to minimize the differences between patients within rooms since it is desirable to combine similar patients. We use age difference as an indicator for the compatibility between patients (see also **?**). Other indicators such as social status, education level, personal background etc. could also be applied in our model with the same logic. In particular, $a_{rt}^{\max} - a_{rt}^{\min}$ denotes the age difference between the oldest and the youngest patient in room r on day t. As such, both auxiliary variables $a_{rt}^{\max}$ and $a_{rt}^{\min}$ are dependent on x_{bp} as well as on the patients already occupying beds. $A_{rt}^{\max}$ ($A_{rt}^{\min}$) is set to the current maximum (minimum) age of all patients already occupying room r on day t. If room r is empty on day t, $A_{rt}^{\max}$ is set to a large integer value that represents the maximum possible age (e.g., 150), and $A_{rt}^{\min}$ is set to 0. Equations (5.7) and (5.8) ensure that the auxiliary variable $a_{rt}^{\max}$ reflects the maximum age of prior occupants and newly allocated patients in a room r on day t. Likewise, Equations (5.9) to (5.11) ensure the same for $a_{rt}^{\min}$ while also making sure that $a_{rt}^{\min}$ equals $a_{rt}^{\max}$ in the event that room r is only occupied by one person or completely empty on day t.

The third term of Equation (5.1) rewards assigning only patients of the same department to specific rooms. Medical rounds for doctors are easier when several patients they are responsible for are in the same room. In addition, walking distances are reduced. Here we need to differentiate between empty and partially occupied rooms. This is indicated by the parameter f_{rt}, which is 1 if room r is empty on day t, and 0 otherwise. Two auxiliary variables y_{rt} and z_{rt} are applied:

- Empty rooms: The auxiliary variable y_{rt} is set to 1 if all patients assigned to an empty room r on day t are from the same medical department, which is achieved by Equations (5.12) and (5.13). Here, d_p is an integer

value that depicts the medical department of patient p and M represents an arbitrary large integer value ("big M").

- Occupied rooms: The auxiliary variable z_{rt} is set to 1 only if all patients assigned to room r are already from the same department. This is achieved by Equations (5.14) to (5.16). Here, D_{rt} is set to 1 if all prior occupants of room r on day t belong to the same medical department, and 0 otherwise.

Finally, the fourth term of the objective function (5.1) is used to balance the workload for nursing staff. This requires the matching of care requirements of patients and care capacity on wards. The specific number of "care units" for every patient p is quantified with c_p. This represents the effort and resources that go into taking care of that particular patient. The available number of nursing staff and thus, workforce or total "care capacity" per ward w and day t is predetermined due to shift schedules, staff rosters, and cannot easily be changed on short notice. Parameter C_{wt} represents the current spare capacity of a given ward w on day t for caring for newly arriving patients (i.e., available capacity, being the delta of the total capacity minus the capacity reserved for current patients in this ward). Exceeding a predefined care capacity per ward w on day t needs to be penalized. The amount by which the capacity of a ward w on day t is exceeded is represented by the auxiliary variable o^+_{wt}. Equations (5.17) and (5.18) link x_{bp} to o^+_{wt}.

5.3.2 Hyper-heuristic

This subsection develops the solution approach. Bed managers require a time-efficient system in everyday work that provides real-time decision support for each new event. An optimal solution approach is impracticable with respect to the combinatorial complexity of the PBA. Other approaches in the literature (see for example Demeester et al. (2010), Ceschia and Schaerf (2011)) also had to resort to using heuristic approaches for the same

reasons. Schäfer et al. (2019) propose a GLA heuristic that derived from the idea of Atkinson (1994). It is able to solve the problem time efficiently, but is vulnerable to ending up in a non-optimal solution. To circumvent these types of situations, we develop a hyper-heuristic framework based on the "pilot method" of Duin and Voß (1999). It supports greedy algorithms in avoiding local optimum traps. Duin and Voß (1999) and Voß et al. (2005) show that the pilot method is suitable for solving highly combinatorial problems (like the PBA), and that it performs competitively compared to well-known meta-heuristics. By only looking forward, the method iteratively weights all options before choosing the most promising. Further notation is delineated in Table 5.3.

a_0	Most promising element $u(a_0) \geq u(a)\ \forall a \in A$
A	Set of all possible choices a, so-called pilots
H	Subheuristic applied to assign remaining pilots $a \in A \setminus S_a$ (e.g., greedy heuristic)
N	Number of partial solutions considered at each iteration
S_a	Partial solution $S_a = a \cup X$
$u(a)$	Predetermined utility function $u : A \to \mathbb{R}$
X	Master solution, iteratively created by adding the most promising element of an iteration $X = X \cup a_0$

Table 5.3: Expanded notation for the pilot method

General Algorithm An initial empty master solution $X = \emptyset$ is iteratively supplemented by an element $a \in A$, whereas A represents the set of all possible choices, so-called pilots. Based on the master solution X, a number of partial solutions N are generated by randomly drawing a pilot ($S_a = a \cup X$). Each partial solution is completed by the remaining pilots $a \in A \setminus S_a$ by applying a subheuristic H. Each solution can be evaluated using a predetermined utility function $u : A \to \mathbb{R}$. Let a_0 be the most promising element $u(a_0) \geq u(a)\ \forall a \in A$. The pilot a_0 gets included in the master solution $X = X \cup a_0$ and excluded from the remaining choices

$A = A \setminus a_0$. Then the algorithm loops to create the next partial solution $S_a = a \cup X$ until a stop criterion is met (e.g., set of pilots is empty $A = \emptyset$, limitation of iterations). In our case, the utility is the total utility of the objective function of Equation (5.1), i.e., $u(a) = U$.

To speed up the computations we limit the solution space by only considering the set of relevant beds $\overline{B}$ and patients $\overline{P}$. The relevant beds considered include only those beds $b, b \in \overline{B}$ that are scheduled to be vacated within the planning horizon T. This means that beds that are already occupied by patients who have an estimated LOS exceeding the planning horizon are not included ($\overline{B} \subseteq B$). Likewise, only those patients $p, p \in \overline{P}, \overline{P} \subseteq P$ who are not yet occupying a bed b within their designated ward space and who require a bed at some point in time within the planning horizon T are considered. In particular, this includes patients who have just arrived, patients who are already waiting in the overflow area, as well as future elective patients already scheduled and anticipated future emergency patients, at some point within the planning horizon T. Limiting the sets for patients and beds is possible, as non-medical room transfers are not allowed. Algorithm 5.1 demonstrates the pilot method tailored to the PBA problem.

Subheuristic The subheuristic applied is based on the GLA heuristic developed by Schäfer et al. (2019). It sequentially calculates the potential added utility value with Equation (5.1) of each possible patient bed combination and also considers at this stage the constraints in Equations (5.2) to (5.19). Finally, it executes the most promising assignment. The additional notation to describe the subheuristic is shown in Table 5.4.

Figure 5.3 illustrates the first iteration of the GLA heuristic. During an initialization process x_{bp} is set to zero and the utility matrix U_{bp} is calculated for all $p \in \overline{P}$ and $b \in \overline{B}$. The utility matrix U_{bp} represents partial utilities that can be added to the total utility function U (Equation (5.1)) by realizing a patient p to bed b assignment.

Algorithm 5.1 Pilot method for PBA

Require: $\overline{P}$, $\overline{B}$, N
Ensure: patient bed assignments x_{bp}
1: $x_{bp} \leftarrow \varnothing$
2: $A \leftarrow$ generatePossiblePatientBedAssignments($\overline{P}, \overline{B}$)
3: **while** ($|A| \neq 0$) **do**
4: **for** $i \leftarrow 1, N$ **do**
5: $a[i] \leftarrow$ random(A)
6: $pilot \leftarrow x_{bp} \cup a[i]$
7: $\overline{B}'[i], \overline{P}'[i] \leftarrow$ updatePatientsAndBeds($\overline{B}, \overline{P}, a[i]$
8: $pilotSolution[i] \leftarrow$ Subheuristic($\overline{B}'[i], \overline{P}'[i]$)
9: $fitness[i] \leftarrow$ calculateFitness($pilotSolution[i]$)
10: **end for**
11: $j \leftarrow$ argmax($fitness$)
12: $a_0 \leftarrow a[j]$
13: $x_{bp} \leftarrow x_{bp} \cup a_0$
14: $\overline{B} \leftarrow \overline{B}'[j]$
15: $\overline{P} \leftarrow \overline{P}'[j]$
16: $A \leftarrow$ updatePossiblePatientBedAssignments(A, a_0)
17: **end while**
18: printPatientBedAssignments(x_{bp})

U_{bp}	Partial utility that an assignment of patient $p, p \in \overline{P}$ to bed b, $b \in \overline{B}$ may add to the total utility U
U_p^{argmax}	Index value of the bed b that adds the maximum partial utility $\max(U_{bp})$ to the total utility U when patient p, $p \in \overline{P}$ is allocated to this bed b
$U_p^{\max}$	Maximum partial utility that assignment of patient p adds to total utility U, $p \in \overline{P}$

Table 5.4: Further notation for the Subheuristic for PBA

If a bed b is not available at any time of the planned stay for the specific patient p, the partial utility value U_{bp} is set to zero. In Iteration I (Step 1), the most promising combination U_{bp} (highest utility value) is chosen, i.e., x_{bp} is set to 1 for patient 2 and bed 6 ($x_{62} = 1$). In the example, patient $p = 2$ is assigned to bed $b = 6$ as this yields the highest partial utility $U_p^{\max}$, with $U_p^{\max} = \max(U_{bp}), \forall b \in \overline{B}, \forall p \in \overline{P}$. To accelerate the process of finding the highest value during the iterations, two auxiliary variables are

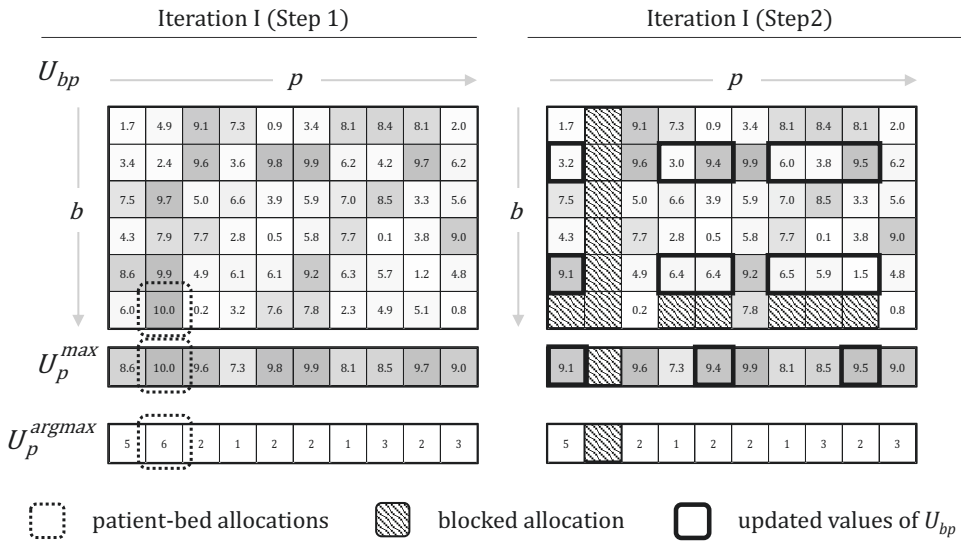


Figure 5.3: Example for the GLA heuristic showing the steps of one iteration

used to indicate the uppermost potential utility of a patient's assignment ($U_p^{\max}$) and the corresponding bed (U_p^{argmax}). This reduces the amount of values that need to be compared from $|\overline{P}| \times |\overline{B}|$ to $|\overline{P}|$ in each step.

The initial allocation of $x_{62} = 1$ has an effect on a series of potential allocation combinations x_{bp} of the remaining patients P and beds B. Subsequently, in Iteration I (Step 2), potential patient bed utilities U_{bp} that have been affected by a previous PBA in Step 1 get updated (black boxes in Figure 5.3). If necessary, $U_p^{\max}$ and U_p^{argmax} are redetermined. The following Iteration II also starts with the assignment of the most beneficial PBA. It will assign patient $p = 6$ to bed $b = 2$, as this has the highest utility U_{bp}, as can be seen on the right of Figure 5.3. In Iteration II, Step 2, the utilities of all remaining patient bed combinations will be updated. This will be continued until all patients are assigned. Algorithm 5.2 represents the iterative, procedural program flow.

Algorithm 5.2 Subheuristic: GLA Heuristic for PBA

Require: $\overline{P}$, $\overline{B}$
Ensure: patient-bed assignments x_{bp}
1: $U_{bp} \leftarrow$ calculatePatientBedMatrix($\overline{P}$, $\overline{B}$)
2: $U_p^{\max} \leftarrow \max(U_{bp})$
3: $U_p^{\text{argmax}} \leftarrow \text{argmax}(U_{bp})$
4: **while** $(\max(U_{bp}) \neq 0)$ **do**
5: $p \leftarrow \text{argmax}(U_p^{\max})$
6: $b \leftarrow U_p^{\max}[p]$
7: $x_{bp} \leftarrow 1$
8: $U_{bp} \leftarrow$ updatePatientBedMatrix(p, b, U_{bp}, $\overline{P}$, $\overline{B}$)
9: **end while**
10: printPatientBedAssignments(x_{bp})

Applied Policies for Patient Bed Assignment To speed up the algorithm and tailor it to the PBA, different policies have been implemented and tested. First, at the start of each new pilot iteration the *filter policy* selects only a determined number of promising pilots. The vector $\text{argmax}(U_p^{\max})$ (see Algorithm 5.2) is used for this, the calculation taking place anyway to subsequently complete the partial solutions. Here, only those pilots with high expected additional utility values are considered. Second, the *drop policy* is applied, which executes the subheuristic H for only a predetermined fraction of the remaining options $a \in A \setminus X$. This can be guaranteed by only considering patients in the subheuristic who arrive within a certain period (shorter than the planning horizon). Finally, we also restricted the *evaluation depth*, i.e., only a subset of pilots $a \subseteq A$ are allocated by the pilot method. The remaining ones $a \in A \setminus X$ get assigned by the subheuristic H. The efficiency and applicability of the different policies are investigated in the numerical studies.

5.4 Numerical study

This section presents numerical studies. We draw upon real-life hospital data from a joint project with a large German hospital. First, we start in subsection 5.4.1 by presenting the data and performing some basic tests. Second, we continue in subsection 5.4.2 by presenting the machine learning approach used to anticipate emergency inpatient arrivals. Third, in subsection 5.4.3 we show the performance of the hyper-heuristic we have developed. Finally, in subsection 5.4.4 we analyze the impact of both the enhanced emergency inpatient arrival forecasting approach as well as the improved hyper-heuristic on the overall solution. All computational steps were carried out using Python 3.8 and R 3.6.

5.4.1 Overview of data

In order to analyze potential influences on emergency patient arrivals, we have gathered metadata on various distinct features that were publicly available and which we suspected of having an impact on the emergency arrivals. These features relate to time and dates, weather data, important local and regional events, as well as historical and current occupancy levels (see Table 5.5). We then used this data in a machine learning approach to anticipate emergency inpatient arrivals based on a selection of the most significant features. The training data used spans across a time period of 2 years from 2014 to 2015, while our test and validation data is taken from 2016.

In a first step, to avoid multicollinearity issues (see e.g., Guyon and Elisseeff (2003)), we determine the Pearson correlation coefficients (PCC) of each potential pairing of features listed in Table 5.5. Figure 5.4 gives an overview of all problematic pairings, i.e., all pairings wherein $|\text{PCC}| >= 0.7$. A simple example of this would be that the maximum temperature T_{max} strongly correlates with the minimum temperature T_{min}, e.g., minimum

Factor	Feature
Time and Date	Weekday (WD_{Mon}, WD_{Tue}, ...)
	Season (Q1, Q2, Q3, Q4)
	School holidays (Hol_{School})
	Bank holidays (Holiday)
	Post holiday weekday ($WD_{postholiday}$)
Weather	Temperature (T_{mean}, T_{min}, T_{max}, T_{dif})
	Air pressure (AP_{mean}, AP_{min}, AP_{max}, AP_{dif})
	Humidity (H_{mean}, H_{min}, H_{max}, H_{dif})
	Wind (W_{mean}, W_{min}, W_{max}, W_{dif}, G_{max})
	Precipitation (Rain, Snow, Hail)
	Snow coverage (S_{cov})
	Storm
Local and Regional Events	Fairs (County Fairs, Sport events)
Current Occupancy	Admissions of previous day (PrevAdmin)

Table 5.5: Overview of factors and properties assessed regarding correlation with emergency inpatient arrivals

and maximum temperatures for any given day during summer time are typically higher than during winter time.

As each medical department is expected to have its its own drivers, we investigate each department individually. The remaining features have to be tested to determine their explanatory power regarding the number of patient arrivals on a given day. This is important for two reasons. First, simply looking at the direct correlation between a given feature and the number of emergency arrivals in the test data can be misleading as this overlooks any potential effects that certain properties only have in combination (Guyon and Elisseeff, 2003). Second, machine learning algorithms tend to be overfitted when the number of features used is significantly higher than optimal (see for example Kohavi and John (1997)). To this end, we make use of the "Boruta" package developed by Kursa and Rudnicki (2010). It consists of a feature selection algorithm based on the "random forest" classification method (Breiman, 2001). Its aim is to rank a set of

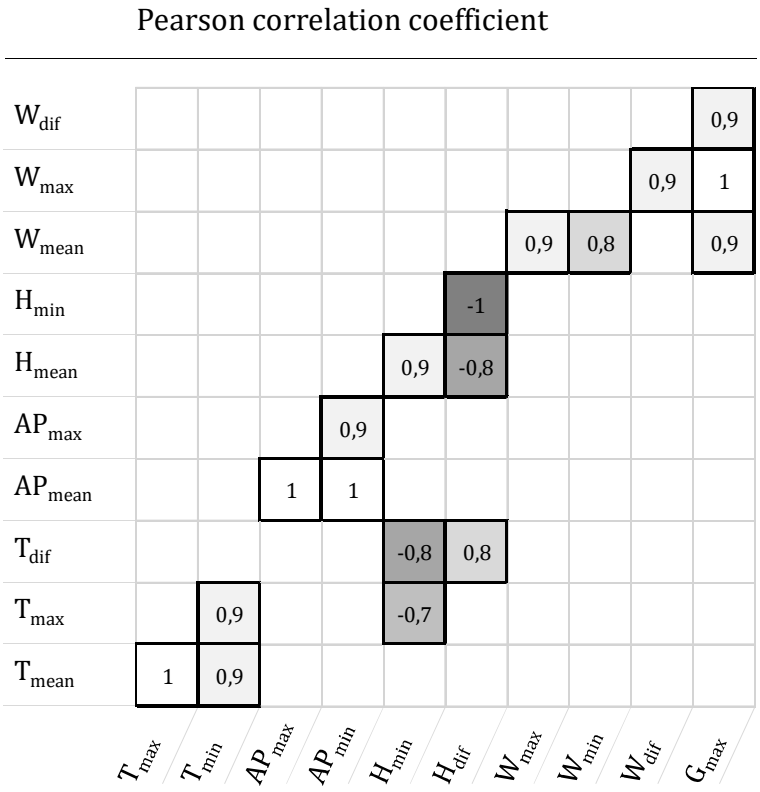
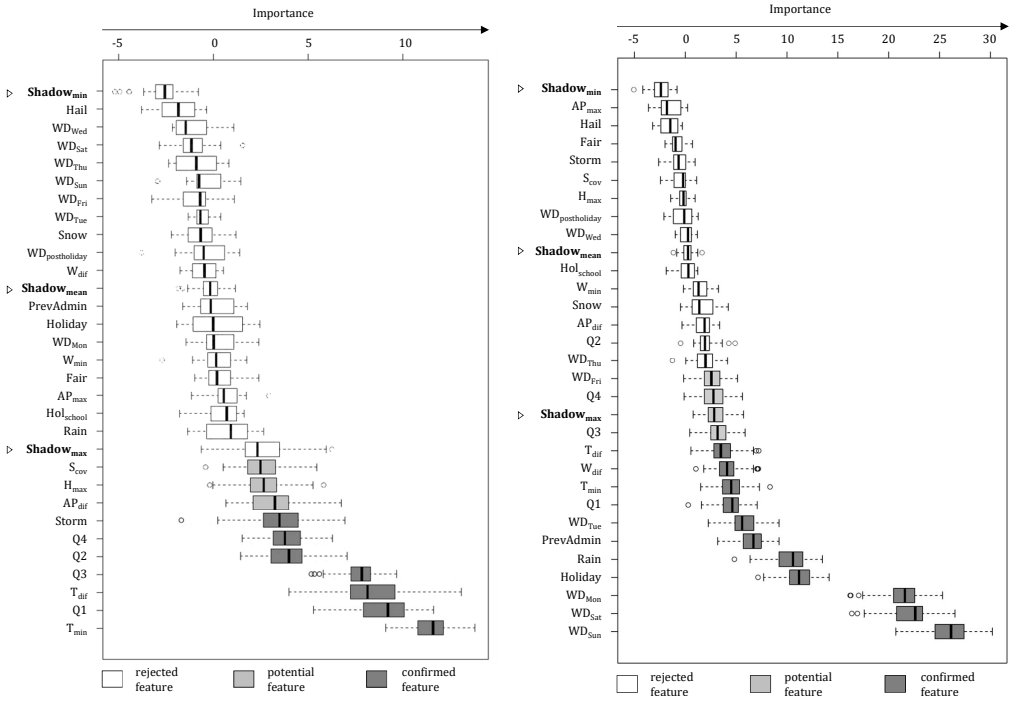


Figure 5.4: Measure of linear correlations between selected parameters

features according to their respective predictive power regarding a specific classification variable, e.g., the number of emergency patient arrivals per day. This ranking is performed according to the individual "importance" of each feature, which is based on the average and standard deviation of the loss of accuracy of classification caused by the random permutation of attribute values between objects. A key idea here is to introduce so-called "shadow variables", i.e., additional random variables, which are then included in the set of existing features. By adding randomness to the data set and collecting results from the ensemble of randomized samples, it is possible to reduce the misleading impact of random fluctuations and correlations.

This process is undergone individually for every medical department, that has emergency arrivals. To give an example, we present detailed results for two different departments, namely trauma surgery and gastroenterology, as can be seen in Figures 5.5a and 5.5b, respectively.



(a) Trauma surgery department

(b) Gastroenterology department

Figure 5.5: Selected outcomes after application of the Boruta package

For trauma surgery, the number of emergency inpatient arrivals is clearly correlated with the seasons (Q1 to Q4), with low temperatures (T_{min}), as well as with the magnitude of intra-day temperature changes (T_{dif}). Naturally, any feature that correlates with the number of emergency inpatient arrivals, in both the training data set and the test data set, can prove useful when anticipating such arrivals. However, the causality behind this correlation may only be guessed. In the case of emergency patients having had an accident that requires trauma surgery, it seems plausible that sudden drops of temperature, which lead to black ice on roads and sidewalks, or typical recreational activities pursued in winter (Q1), e.g., skiing, are responsible for this effect.

For the gastroenterology department, however, the picture looks quite different. Here, holidays, weekends and Mondays each exhibit a high explanatory correlation with regard to incoming emergency patients, whereas the temperature has a considerably lower influence when compared to the trauma surgery department. This could be due to a couple of different reasons. For instance, doctors and nursing staff we interviewed have reported that many gastroenterological illnesses often initially present with non-specific abdominal pain symptoms, which then intensify over the course of several days. This means that in comparison with a broken hip, for example, there is no immediate need to get to a hospital, such that patients could opt to stay home on weekends. An alternative explanation could be that resident doctors' offices are typically closed on weekends and patients who are not yet aware of the severity of their illness will usually wait until the next workday to see their family doctor who might then immediately refer them to a hospital for further diagnosis and treatment.

To summarize, the drivers for the arrival of emergency patients are different across departments. This requires to address the forecasting and PBA problem by department.

5.4.2 Applying machine learning to estimate emergency patients

Estimating the number of future emergency patient admissions is inherently a regression problem. We therefore first applied (1) regression-based methods using the metadata described in Table (5.5). In addition, in a (2) second step we applied a multilayer artificial neural network to account for nonlinear dependencies. We used regularization methods in both approaches to avoid overfitting. Finally, (3) we used the test data to evaluate the generalization abilities of our trained models.

(1) Regression-based methods Ridge regression (RR) uses l_2-regularization (Hoerl and Kennard, 1970), whereas LASSO (LR) uses l_1-regularization (Tibshirani, 1996). l_2-regularization accounts for correlations between the input features, while l_1-regularization favors sparse solutions. Elastic Net (EN) is a regression-based method that combines l_1 and l_2 regularization (Zou and Hastie, 2005). Another class of regression models is Group-LASSO (GL), which allows individual features to be combined into groups (Yuan and Lin, 2006). All features of a group are penalized together, leading to whole groups being considered or neglected. We used 10-fold cross-validation to tune the hyperparameter λ for each approach, which controls the strength of the regularization. For EN we performed a grid-search between 0 and 1 in 0.025 steps to optimize the hyperparemters λ_1 and λ_2, which are used to control the l_1 and l_2 penalty respectively.

(2) Artificial neural network We used artificial neural networks (ANN) (Goodfellow et al., 2016; LeCun et al., 2015) to account for non-linear dependencies. An example architecture of an ANN is illustrated in Figure (5.6). We have evaluated several typologies of ANNs by varying the number of hidden layers between one to five. The best results have been achieved by applying a "32:16:8:4:2" network (the numbers are the number of neurons per hidden layer; hidden layers are separated by colons), the rectified linear unit (ReLu) as activation function, l_1 and l_2 regularization and the mean-squared error (MSE) loss function as well as the optimizer RMSprop. To avoid overfitting we have investigated the learning curve of training and validation loss. For tuning hyperparameters l_1 and l_2 we used a grid search algorithm.

(3) Evaluation of performance on test data We applied the learned models to the test data from four departments at our case hospital that have a significant number of emergency patients. For example, orthopedics has almost no emergency patients. Table 5.6 summarizes the results and shows the root mean square error (RMSE), the machine learning model used

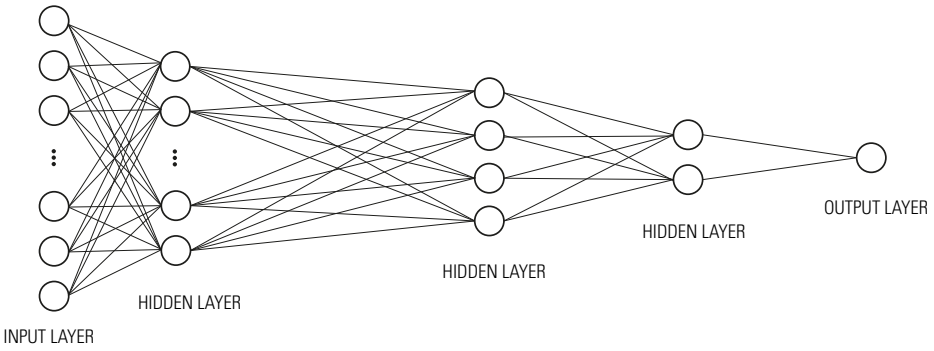


Figure 5.6: Example of the structure of a neural network, including three hidden layers

	Approach 1	Approach 2 (ML)					Max.	
		RR	LR	EN	GL	ANN	Improvement	
Department	RMSE	RMSE	RMSE	RMSE	RMSE	RMSE	[%]	Type
Department 1	4.888	4.331	4.309	4.183	4.103	**4.060**	16.9	ANN
Department 2	4.108	3.892	3.847	3.848	**3.675**	3.835	10.5	GL
Department 3	2.888	2.887	2.824	2.8	**2.743**	2.778	5.021	GL
Department 4	3.535	3.126	3.099	3.097	**3.067**	3.192	13.2	GL

Table 5.6: Anticipation of emergency inpatient arrivals using machine learning

that achieved the best performance, as well as the improvement achieved in comparison to historical averages. The historical averages serve as a baseline approach, and this is denoted as "Approach 1". This is compared with our above-described machine learning approach (denoted as "Approach 2 (ML)"). Table 5.6 shows that the machine learning approach outperforms "Approach 1". The ML approach leads to improvements of up to 17%, depending on the department, compared to the basic historical averages.

5.4.3 Performance of the hyper-heuristic

In order to assess the solution quality of the hyper-heuristic proposed in this paper, we drew upon nine data sets. The available real data of

one department cluster consisting of department 1 and department 2 (see also 5.4.2) based on actual patient movements between January 2016 and September 2016 will be considered. The department cluster consists of six wards with 24 beds each. Each data set is composed of 28 consecutive days and comprises an average of 648 unique patients. On average, 40% of patients are men and 60% women with an average age of 70 years and a length of stay of 6 days. All data sets reveal high ratios of emergency patients, e.g., up to 90%. We set the parameters in alignment with currently applied weights in our case hospital: $\alpha = 1$, $\beta = 0.1$, $\gamma, \delta = 2$, $q = 0.01$. Furthermore, the weighting factor Ξ_p was set to three distinct values depending on the patient type. Notably, these consist of $\Xi^{el} = 10$ for elective patients, $\Xi^{em} = 9$ for current emergency arrivals, and $\Xi^{an} = 4$ for anticipated emergency arrivals. Here, elective patients are preferred to current emergency patients, and these in turn are given preference vs. expected future emergency arrivals. We have adjusted the existing data by eliminating all uncertainty factors for the sole purpose of monitoring the performance of the heuristics applied. Accordingly, emergency patients treated like elective patients and their exact admission are known in advance. Both patient types are no longer subject to LOS updates due to precisely known discharge times. Furthermore, patient no-shows are neglected. This means that the data sets considered are no longer affected by stochastic variations and are assumed to be deterministic.

Application to single problem instances We first assessed the performance of our hyper-heuristic for single problem instances. Using such a static version is a usual benchmark approach (see literature review in Section 5.2.3.1 above and for example in Bilgin et al. (2012); Guido et al. (2018); Dorgham et al. (2019); Ceschia and Schaerf (2011)). This approach excludes parameter-dependent (e.g., planning horizon, time-dependent relevance) performance differences caused by time series analysis. These parameters could lead to worse performance in the time series analysis and thus reduce the meaningfulness of hyper-heuristic performance despite better performance in all single problem instances. We tested several policies (see Section

5.3.2). In particular, we applied a *filter policy* with which we restricted the number of promising pilots to different predetermined amounts, which were determined based on their individual additional potential benefit to the utility function prior to an algorithm run-through. The best patient-bed assignments are drawn randomly from the five most promising patients. This was done to avoid unnecessary computational effort while at the same time ensuring that no potentially "lucrative" PBAs are overlooked. It should be noted that several potential PBAs of a single patient may have similar values and hence a wide variety of alternative promising PBAs exist. In addition, we applied a *drop policy* by limiting the application of the GLA subheuristic to only those patients that were known or anticipated to arrive within a certain number of days, which also leads to a significant reduction of computational time while retaining a high solution quality. Finally, we varied the *evaluation depth* by restricting the amount of subsequent PBAs obtained through the pilot method. To give an example, selecting only 10 pilots and a depth of 20 translates into applying the pilot method to determine the first 20 PBA, wherein for each of these 20 assignments the 10 most promising pilots will be chosen and evaluated using the GLA heuristic. Table 5.7 gives an overview of the solutions obtained. For each of the shown combinations of data set used, amount of promising pilots filtered (in lines), and evaluation depth (in columns), we have taken into account all single problem instances which emerged by executing the data sets. This results in around 2,000 single problem instances for each data set (i.e., around 288,000 in total), promising pilot and evaluation depth combinations. We did this in order to account for statistical distributions, which arise due to the inherent randomness associated with our implementation of the hybrid-heuristic.

The results obtained allow for drawing three main insights. First, by using the pilot method, it was possible to increase the solution quality in comparison to the GLA heuristic by up to 2.90% while achieving an average increase of 2.42% when considering 20 promising pilots combined with an evaluation depth of 20. This number can of course vary depending on the characteristics of the underlying patient clientele. However, the

effect observed is substantially the same across all nine data sets. Second, as is to be expected, increasing the evaluation depth as well as increasing the number of promising pilots both lead to an increase in solution quality. This is because it is more likely that better solutions will be found when broadening the search space as this increases the chance of finding solutions that are further away from standard GLA heuristic solutions. The effect of increasing the evaluation depth has a higher impact on solution quality than increasing the number of promising pilots considered. A reason for this effect could be seen in that even when using a low number of promising pilots considered, the pilots chosen exhibit the highest additional benefit to the overall utility function, respectively, which makes the underlying PBA more likely to be part of a good solution. Third, depending on the situation at hand, the acquired gain in solution quality due to a broader search space goes hand in hand with higher computational effort, which can be an important factor when requiring real-time PBAs in actual applications. Roughly speaking, the total computation time for a single problem instance can be estimated by adding up the total number of times the subheuristic has to run through all PBAs for a given single problem instance. To give an example, an evaluation depth of 10 combined with 10 selected pilots will add up to 100 applications of the subheuristic while an evaluation depth of 5 combined with 5 selected pilots will only require 25 run-throughs of the GLA heuristic, or 25% of the time. The runtime changes only proportional to the dimension of evaluation depth when multi-processing is applied. This means that the runtime compared to the GLA heuristic is just multiplied by the evaluation depth. The GLA heuristic is typically solved in an average of less than one second for instances encompassing 124 beds.

Application to time series In addition to comparing the solution quality for single problem instances, we have undertaken analyses to compare the performances of both approaches over time. For this purpose, the data sets that have been cleared of uncertainties are also used. Furthermore, to investigate the scaling effect in relation to the department cluster size we divided the nine existing data sets with regard to the department cluster

size stepwise by 24 beds from 24 to 120. To do this, the patients and beds are added depending on the division of the wards and their specific specialty. To test the hyper-heuristic approach developed, we use the top-performing settings from the single problem instance analyses (see Table 5.7), i.e., an evaluation depth of 20 combined with a selection of 20 promising pilots for each subsequent PBA.

The results of this analysis are presented in Table 5.8. Again, we have accounted for statistical effects of the stochastic search procedure by running the algorithm 20 times for each combination of data set and beds considered. Here, the results show an increase in total utility. The hyper-heuristic approach outperforms the GLA heuristic by 1.48% on average while achieving an increase of up to 3.86% for certain data sets. The utility increase of the hyper-heuristic vs. the GLA heuristic for the time series analyses in Table 5.8 is not as clearly predictable as for the single problem instance solution in Table 5.7. This is due to the settings of the hyper-heuristic (i.e., planning horizon, time-dependent relevance parameter Q_t). Furthermore, only patients within the planning horizon, that may overlap with the hospital stays of future elective patients (arrival exceeding planning horizon) are considered. In other words, even if the hyper-heuristic performs considerably better than the GLA heuristic for each single problem instance within the time series investigated, time-dependent parameter settings may eradicate the positive effect of the hyper-heuristic compared to the GLA heuristic for certain combinations of data sets and beds. This also explains some negative entries in the minimum values of Table 5.8. The hyper-heuristic outperformed the GLA heuristic in over 99% of the test instances.

5.4.4 Hyper-heuristic combined with enhanced emergency inpatient arrival forecasting

In this subsection, the impact of both the enhanced emergency inpatient arrival forecasting approach as well as the improved hyper-heuristic with

regard to real data are analyzed. The nine data sets (6 wards with 24 beds each), including uncertainties, are used to do this. Each data set consists of around 2,000 unique events that take place over the course of 28 days. We denote the hyper-heuristic approach including the enhanced emergency patient admission data, which was achieved with machine learning, as *Hyper-Heuristic ML*. It is executed 20 times for all data sets and the average of all runs is reported. We apply two benchmarks:

- GLA Avg: The first is the GLA heuristic of Schäfer et al. (2019) where the arrivals of emergency patients have been estimated according to Approach 1 (see 5.4.2). We normalize all values of the alternative approaches to this.
- GLA ML: The second is also based on the GLA heuristic, but the arrivals of emergency patients have also been estimated with machine learning.

Looking at the results of the analysis of the three methods in Table 5.9, the normalized values of the objective function give a first indication of the performance of our approach. It can be noted that the Hyper-Heuristic ML outperforms the GLA as well as the GLA ML approach in each data set. On average across all data sets the Hyper-Heuristic ML shows 1.4% better results than the GLA and beats the GLA ML by 0.4%. Even the minimum outcome of the Hyper-Heuristic ML for all data sets performs better than the GLA method. This makes the Hyper-Heuristic ML the most promising and reliable approach to solve the PBA problem.

5.5 Conclusion and further areas of research

Conclusion This paper develops and investigates improvements for the operational PBA. The model used has been developed in a joint project with a large German hospital covering all major disciplines and incorporates the objectives and constraints of the three main stakeholders, namely patients, doctors, and nursing staff. It integrates the planning of current emergency

and elective patient arrivals, future elective patient arrivals, as well as anticipated future emergency patient arrivals. Two important aspects were tackled and improved in this paper.

- To tackle the uncertainty of emergency patient admissions, we applied machine learning techniques to estimate these more precisely. To this end, we used historic emergency inpatient data as well as metadata relating to time, date, weather forecasts, and local and regional events. We are the first to investigate and make use of the correlation of several external factors, such as weather data, to better anticipate emergency inpatient admissions.
- To enhance the performance of the solution approach we integrate the GLA heuristic into the Pilot method which consists of a hyper-heuristic framework.

Our numerical results have shown that machine learning approaches can outperform historical average approaches by up to 17% when it comes to predicting emergency inpatient arrivals. The underlying drivers for emergency inpatient arrivals differ strongly between departments due to the associated patient clientele, e.g., Trauma Surgery shows a higher dependency on weather data than Gastroenterology, which in turn is more strongly correlated with times and dates. Compared to the GLA heuristic, the hyper-heuristic developed can improve performance by up to 3% for single problem instances and up to 4% in a time series analysis. With respect to real data, the hyper-heuristic approach combined with sophisticated prediction of future emergency patient admissions by machine learning outperforms the GLA heuristic in a time series analysis by up to 2.2%.

Future areas of research Various opportunities exist for further research. For the problem shown, the existing solution methods can be further developed and different approaches can be pursued. The focus may be on enhanced anticipation of the input parameters, improvement of the heuristic methods or development of an optimal solution method.

The estimate of input parameters focuses on both emergency and elective patients. Information on the progress the patient's recovery is making (e.g., LOS as well as type and probability of complications) can be anticipated for both patient groups. The approximation of time-related arrivals and patient characteristics (e.g., gender, age, and disease) is especially in focus for emergency patients, while no-show rates are interesting for elective patients. In the development of heuristics, the focus can be on runtime-related aspects, solution quality or the proportion of both by implementing and testing alternative approaches (e.g., meta-heuristics, matheuristics, exact approaches). Another topic of research interest is to integrate upstream and/or downstream processes in the decision model, such as admission scheduling of elective patients, operating room scheduling, bed transport services or staff rostering (cf. e.g., van Oostrum et al. (2008); Beaudry et al. (2010); Rachuba and Werners (2014); Aringhieri et al. (2015); Erhard et al. (2018); Thielen (2018); Séguin et al. (2019)). This integration makes it possible to obtain information about conflicts of interests of individual problems. In order to maximize profit, operating rooms should usually be booked to full capacity, although the hospital may not have suitable beds available for patients who have had surgery. Furthermore, the underlying mechanics of the PBA decision model are not limited to hospital settings alone. Further investigation could be made into identifying problem settings that have a similar scope. To give an example, the student-room assignment problem in hostels (Alfred and Yu, 2020) could potentially yield further areas of application.

DS[1]1	**Depth**					DS 2	**Depth**				
Pilots[2]	5	10	15	20	Avg.	**Pilots**	5	10	15	20	Avg.
5	0.66%	1.13%	1.49%	1.73%	1.25%	5	0.52%	0.99%	1.33%	1.49%	1.08%
10	0.76%	1.27%	1.60%	1.91%	1.38%	10	0.61%	1.12%	1.45%	1.62%	1.20%
15	0.79%	1.30%	1.59%	1.94%	1.40%	15	0.65%	1.17%	1.46%	1.63%	1.22%
20	0.83%	1.31%	1.66%	1.93%	1.43%	20	0.64%	1.21%	1.41%	1.61%	1.21%
Avg.	0.76%	1.25%	1.59%	1.88%	-	Avg.	0.60%	1.12%	1.41%	1.59%	-
DS 3	**Depth**					DS 4	**Depth**				
Pilots	5	10	15	20	Avg.	**Pilots**	5	10	15	20	Avg.
5	0.55%	0.98%	1.49%	2.05%	1.02%	5	0.65%	1.31%	1.98%	2.39%	1.56%
10	0.60%	1.06%	1.6%	2.27%	1.10%	10	0.78%	1.64%	2.15%	2.57%	1.76%
15	0.66%	1.18%	1.79%	2.20%	1.16%	15	0.81%	1.62%	2.35%	2.78%	1.85%
20	0.68%	1.28%	1.92%	2.38%	1.24%	20	0.79%	1.77%	2.28%	2.67%	1.85%
Avg.	0.62%	1.12%	1.70%	2.22%	-	Avg.	0.76%	1.58%	2.19%	2.60%	-
DS 5	**Depth**					DS 6	**Depth**				
Pilots	5	10	15	20	Avg.	**Pilots**	5	10	15	20	Avg.
5	0.68%	1.12%	1.51%	1.83%	1.28%	5	0.81%	1.51%	2.27%	2.54%	1.75%
10	0.78%	1.21%	1.64%	2.04%	1.41%	10	0.91%	1.73%	2.44%	2.85%	1.94%
15	0.79%	1.26%	1.78%	1.98%	1.44%	15	0.96%	1.82%	2.47%	2.79%	1.97%
20	0.84%	1.33%	1.80%	2.03%	1.49%	20	0.95%	1.91%	2.51%	2.90%	2.02%
Avg.	0.77%	1.23%	1.68%	1.97%	-	Avg.	0.91%	1.74%	2.42%	2.77%	-
DS 7	**Depth**					DS 8	**Depth**				
Pilots	5	10	15	20	Avg.	**Pilots**	5	10	15	20	Avg.
5	0.82%	1.43%	2.02%	2.40%	1.65%	5	0.95%	1.68%	2.22%	2.66%	1.86%
10	0.88%	1.64%	2.17%	2.62%	1.81%	10	1.06%	1.83%	2.43%	2.72%	1.99%
15	0.93%	1.76%	2.31%	2.70%	1.91%	15	1.11%	1.87%	2.51%	2.69%	2.03%
20	0.94%	1.77%	2.32%	2.79%	1.94%	20	1.12%	1.92%	2.52%	2.84%	2.08%
Avg.	0.89%	1.65%	2.21%	2.63%	-	Avg.	1.06%	1.82%	2.42%	2.73%	-
DS 9	**Depth**					Total[3]	**Depth**				
Pilots	5	10	15	20	Avg.	**Pilots**	5	10	15	20	Avg.
5	0.70%	1.19%	1.72%	2.33%	1.46%	5	0.70%	1.26%	1.78%	2.16%	1.48%
10	0.78%	1.43%	1.98%	2.55%	1.66%	10	0.80%	1.44%	1.94%	2.35%	1.63%
15	0.80%	1.47%	2.06%	2.52%	1.69%	15	0.83%	1.49%	2.03%	2.36%	1.68%
20	0.83%	1.52%	2.16%	2.63%	1.75%	20	0.84%	1.56%	2.07%	2.42%	1.72%
Avg.	0.78%	1.40%	1.98%	2.51%	-	Avg.	0.79%	1.44%	1.96%	2.32%	-

[1] Data set used to extract problem instances

[2] Number of promising pilots filtered for further analysis

[3] Total average across all problem instances analyzed

Table 5.7: Solution quality of the Pilot method compared to the GLA heuristic for single problem instances

Data Set	24 beds			Data Set	48 beds		
	Min	Avg.	Max		Min	Avg.	Max
1	0.64%	1.62%	3.4%	1	1.54%	2.47%	3.58%
2	0.43%	0.75%	1.32%	2	- 0.02%	1.69%	2.73%
3	- 1.12%	- 0.03%	0.68%	3	0.72%	0.99%	1.34%
4	1.08%	1.31%	1.54%	4	1.87%	3.18%	3.71%
5	0.14%	1.06%	2.60%	5	1.27%	2.06%	2.55%
6	1.89%	2.83%	3.48%	6	1.59%	1.81%	2.16%
7	1.66%	1.77%	1.96%	7	1.85%	2.56%	3.86%
8	0.82%	1.72%	2.44%	8	1.18%	2.87%	3.68%
9	- 0.45%	0.46%	1.50%	9	1.21%	2.05%	2.89%
Avg.	0.57%	1.28%	2.10%	Avg.	1.25%	2.19%	2.95%
Data Set	**72 beds**			**Data Set**	**96 beds**		
	Min	Avg.	Max		Min	Avg.	Max
1	0.79%	1.62%	2.39%	1	1.00%	1.69%	2.17%
2	1.32%	2.30%	3.18%	2	0.98%	1.55%	1.88%
3	1.47%	1.47%	1.47%	3	0.00%	0.23%	0.47%
4	0.67%	0.67%	0.67%	4	0.81%	1.17%	1.61%
5	1.75%	1.75%	1.75%	5	1.36%	1.65%	1.82%
6	0.68%	0.68%	0.68%	6	0.80%	1.68%	2.38%
7	2.25%	2.25%	2.25%	7	1.53%	1.71%	1.97%
8	1.52%	1.52%	1.52%	8	0.32%	0.84%	1.37%
9	1.02%	1.02%	1.02%	9	0.17%	0.58%	0.76%
Avg.	1.27%	1.47%	1.66%	Avg.	0.77%	1.23%	1.61%
Data Set	**120 beds**			**Data Set**	**Total[1]**		
	Min	Avg.	Max		Min	Avg.	Max
1	1.30%	1.75%	2.11%	1	1.06%	1.83%	2.73%
2	0.88%	1.24%	1.57%	2	0.72%	1.51%	2.13%
3	1.03%	1.61%	2.06%	3	0.42%	0.85%	1.21%
4	- 0.14%	0.56%	1.03%	4	0.86%	1.38%	1.71%
5	0.69%	0.91%	1.20%	5	1.04%	1.49%	1.98%
6	1.92%	2.30%	2.74%	6	1.38%	1.86%	2.29%
7	1.18%	1.48%	1.86%	7	1.69%	1.95%	2.38%
8	- 0.05%	0.25%	0.52%	8	0.76%	1.44%	1.91%
9	0.76%	1.07%	1.42%	9	0.55%	1.04%	1.52%
Avg.	0.84%	1.24%	1.61%	Avg.	0.94%	1.48%	1.98%

[1] Total average across all bed sizes

Table 5.8: Solution quality of the Pilot method compared to the GLA heuristic for time series analysis

Model	**Data Set**				
	1	2	3	4	5
GLA Avg	99.37%	102.13%	102.77%	107.8%	99.65%
GLA ML	100.12%	102.22%	103.69%	107.83%	101.60%
Hyper-Heuristic ML	100.38%	102.74%	104.31%	108.25%	101.77%

Model	**Data Set**				**Total**[1]
	6	7	8	9	
GLA Avg.	98.32%	101.11%	95.05%	93.81%	100.00%
GLA ML	100.39%	101.96%	96.27%	94.91%	101.00%
Hyper-Heuristic ML	100.46%	102.21%	97.12%	95.31%	101.40%

[1] total average across all data sets

Table 5.9: Solution quality of the Hyper-Heuristic ML compared to benchmarks using a time series analysis

Bibliography

Afilal, M., Yalaoui, F., Dugardin, F., Amodeo, L., Laplanche, D., Blua, P., 2016. Forecasting the emergency department patients flow. Journal of Medical Systems 40.

Alfred, R., Yu, H.F., 2020. Automated scheduling of hostel room allocation using genetic algorithm, in: Sharma, N., Chakrabarti, A., Balas, V.E. (Eds.), Data Management, Analytics and Innovation, Springer, Singapore. pp. 151–160.

Aringhieri, R., Landa, P., Soriano, P., Tànfani, E., Testi, A., 2015. A two level metaheuristic for the operating room scheduling and assignment problem. Computers & Operations Research 54, 21–34.

Atkinson, J.B., 1994. A greedy look-ahead heuristic for combinatorial optimization: An application to vehicle scheduling with time windows. The Journal of the Operational Research Society 45, 673–684.

Bastos, L.S., Marchesi, J.F., Hamacher, S., Fleck, J.L., 2019. A mixed integer programming approach to the patient admission scheduling problem. European Journal of Operational Research 273, 831–840.

Beaudry, A., Laporte, G., Melo, T., Nickel, S., 2010. Dynamic transportation of patients in hospitals. OR Spectrum 32, 77–107.

Bekker, R., Koole, G., Roubos, D., 2016. Flexible bed allocations for hospital wards. Health Care Management Science 20, 453–466.

Beliën, J., Demeulemeester, E., 2007. Building cyclic master surgery schedules with leveled resulting bed occupancy. European Journal of Operational Research 176, 1185–1204.

Best, T.J., Sandıkçı, B., Eisenstein, D.D., Meltzer, D.O., 2015. Managing hospital inpatient bed capacity through partitioning care into focused wings. Manufacturing & Service Operations Management 17, 157–176.

Bilgin, B., Demeester, P., Misir, M., De Vancroonenburg, W., Vanden Berghe, G., 2012. One hyper-heuristic approach to two timetabling problems in health care. Journal of Heuristics 18, 401–434.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Bron, C., Kerbosch, J., 1973. Finding all cliques of an undirected graph. Communications of the ACM 16, 575–579.

Carvalho-Silva, M., Monteiro, M.T.T., de Sá-Soares, F., Dória-Nóbrega, S., 2018. Assessment of forecasting models for patients arrival at emergency department. Operations Research for Health Care 18, 112–118.

Ceschia, S., Schaerf, A., 2011. Local search and lower bounds for the patient admission scheduling problem. Computers & Operations Research 38, 1452–1463.

Ceschia, S., Schaerf, A., 2012. Modeling and solving the dynamic patient admission scheduling problem under uncertainty. Artificial intelligence in medicine 56, 199–205.

Ceschia, S., Schaerf, A., 2016. Dynamic patient admission scheduling with operating room constraints, flexible horizons, and patient delays. Journal of Scheduling 19, 377–389.

Cochran, J.K., Bharti, A., 2006. Stochastic bed balancing of an obstetrics hospital. Health Care Management Science 9, 31–45.

Cochran, J.K., Roche, K., 2008. A queuing-based decision support methodology to estimate hospital inpatient bed demand. Journal of the Operational Research Society 59, 1471–1482.

Cotta, C., 2011. Effective patient prioritization in mass casualty incidents using hyperheuristics and the pilot method. OR Spectrum 33, 699–720.

De Bruin, A., Bekker, R., Van Zanten, L., Koole, G.M., 2010. Dimensioning hospital wards using the erlang loss model. Annals of Operations Research 178, 23–43.

Demeester, P., Souffriau, W., Causmaecker, P.d., Berghe, G.V.d., 2010. A hybrid tabu search algorithm for automatically assigning patients to beds. Artificial Intelligence in Medicine 48, 61–70.

Demeulemeester, E., Beliën, J., Cardoen, B., Samudra, M., 2013. Operating room planning and scheduling. Handbook of Healthcare Operations Management .

Dorgham, K., Nouaouri, I., Ben-Romdhane, H., Krichen, S., 2019. A hybrid simulated annealing approach for the patient bed assignment problem. Procedia Computer Science 159, 408–417. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 23rd International Conference KES2019.

Duin, C., Voß, S., 1999. The pilot method: A strategy for heuristic repetition with application to the steiner problem in graphs. Networks 34, 181–191.

Erhard, M., Schoenfelder, J., Fügener, A., Brunner, J.O., 2018. State of the art in physician scheduling. European Journal of Operational Research 265, 1–18.

Fügener, A., Hans, E.W., Kolisch, R., Kortbeek, N., Vanberkel, P.T., 2014. Master surgery scheduling with consideration of multiple downstream units. European Journal of Operational Research 239, 227–236.

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

Gartner, D., Padman, R., 2019. Flexible hospital-wide elective patient scheduling. Journal of the Operational Research Society 0, 1–15.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Green, L.V., Ngyuen, V., 2001. Strategies for cutting hospital beds: the impact on patient service. Health Service Research 36, 421–442.

Gross, C.N., Fügener, A., Brunner, J., 2017. Online rescheduling of physicians in hospitals. Flexible Services and Manufacturing Journal 7, 1–33.

Guido, R., Groccia, M.C., Conforti, D., 2018. An efficient matheuristic for offline patient-to-bed assignment problems. European Journal of Operational Research 268, 486–503.

Gul, M., Celik, E., 2018. An exhaustive review and analysis on applications of statistical forecasting in hospital emergency departments. Health Systems , 1–22.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of Maching Learning Research 3, 1157–1182.

Hall, R. (Ed.), 2012. Handbook of Healthcare System Scheduling. International Series in Operations Research & Management Science, Springer US.

Harris, R.A., 1986. Hospital bed requirements planning. European Journal of Operational Research 25, 121–126.

Helber, S., Böhme, D., Oucherif, F., Lagershausen, S., Kasper, S., 2015. A hierarchical facility layout planning approach for large and complex hospitals. Flexible Services and Manufacturing Journal , 1–25.

Herring, W.L., Herrmann, J.W., 2012. The single-day surgery scheduling problem: sequential decision-making and threshold-based heuristics. OR Spectrum 34, 429–459.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Hof, S., Fügener, A., Schoenfelder, J., Brunner, J., 2015. Case mix planning in hospitals: a review and future agenda. Health Care Management Science , 1–14.

Holm, L.B., Lurås, H., Dahl, F.A., 2013. Improving hospital bed utilisation through simulation and optimisation: With application to a 40% increase in patient volume in a norwegian general hospital. International Journal of Medical Informatics 82, 80–89.

Hübner, A., Kuhn, H., Walther, M., 2018. Combining clinical departments and wards in maximum-care hospitals. OR Spectrum 40, 679–709.

Hübner, A., Walther, M., Kuhn, H., 2016. Approach to clustering clinical departments, in: Matta, A., Sahin, E., Li, J., Guinet, A., Vandaele, N.J. (Eds.), Health Care Systems Engineering for Scientists and Practitioners. Springer Proceedings in Mathematics & Statistics, Lyon, pp. 111–120.

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

Hulshof, P.J.H., Mes, M.R.K., Boucherie, R.J., Hans, E.W., 2016. Patient admission planning using approximate dynamic programming. Flexible Services and Manufacturing Journal 28, 30–61.

Kifah, S., Abdullah, S., 2015. An adaptive non-linear great deluge algorithm for the patient-admission problem. Information Sciences 295, 573–585.

Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artificial Intelligence 97, 273–324.

Kursa, M., Rudnicki, W., 2010. Feature selection with the boruta package. Journal of Statistical Software, Articles 36, 1–13.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

Lusby, R.M., Schwierz, M., Range, T.M., Larsen, J., 2016. An adaptive large neighborhood search procedure applied to the dynamic patient admission scheduling problem. Artificial Intelligence in Medicine 74, 21–31.

Luscombe, R., Kozan, E., 2016. Dynamic resource allocation to improve emergency department efficiency in real time. European Journal of Operational Research 255, 593–603.

van Oostrum, J.M., Van Houdenhoven, M., Hurink, J.L., Hans, E.W., Wullink, G., Kazemier, G., 2008. A master surgical scheduling approach for cyclic scheduling in operating room departments. OR Spectrum 30, 355–374.

Prim, R.C., 1957. Shortest connection networks and some generalizations. The Bell System Technical Journal 36, 1389–1401.

Rachuba, S., Werners, B., 2014. A robust approach for scheduling in hospitals using multiple objectives. Journal of the Operational Research Society 65, 546–556.

Range, T.M., Lusby, R.M., Larsen, J., 2014. A column generation approach for solving the patient admission scheduling problem. European Journal of Operational Research 235, 252–264.

Schäfer, F., Walther, M., Hübner, A., Kuhn, H., 2019. Operational patient-bed assignment problem in large hospital settings including overflow and uncertainty management. Flexible Services and Manufacturing Journal 31, 1012–1041.

Schiele, J., Koperna, T., Brunner, J.O., 2019. Predicting bed occupancy for integrated surgery scheduling via neural networks. Working Paper University of Augsburg.

Schmidt, R., Geisler, S., Spreckelsen, C., 2013. Decision support for hospital bed management using adaptable individual length of stay estimations and shared resources. BMC Medical Informatics and Decision Making 13, 1–19.

Séguin, S., Villeneuve, Y., Blouin-Delisle, C.H., 2019. Improving patient transportation in hospitals using a mixed-integer programming model. Operations Research for Health Care 23, 100202.

Taramasco, C., Olivares, R., Munoz, R., Soto, R., Villar, M., de Albuquerque, V.H.C., 2019. The patient bed assignment problem solved by autonomous bat algorithm. Applied Soft Computing 81, 105484.

Thielen, C., 2018. Duty rostering for physicians at a department of orthopedics and trauma surgery. Operations Research for Health Care 19, 80–91.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58, 267–288.

Turhan, A.M., Bilgen, B., 2017. Mixed integer programming based heuristics for the patient admission scheduling problem. Computers & Operations Research 80, 38–49.

Van Essen, T.J., Van Houdenhoven, M., Hurink, J.L., 2015. Clustering clinical departments for wards to achieve a prespecified blocking probability. OR Spectrum 37, 243–271.

Vanberkel, P.T., Boucherie, R.J., Hans, E.W., Hurink, J.L., Litvak, N., 2012. Efficiency evaluation for pooling resources in health care. OR Spectrum 34, 371–390.

Vancroonenburg, W., De Causmaecker, P., Vanden Berghe, G., 2016. A study of decision support models for online patient-to-room assignment planning. Annals of Operations Research 239, 253–271.

Vassilacopoulos, G., 1985. A simulation model for bed allocation to hospital inpatient departments. SIMULATION 45, 233–241.

Voß, S., Fink, A., Duin, C., 2005. Looking ahead with the pilot method. Annals of Operations Research 136, 285–302.

Wargon, M., Guidet, B., Hoang, T.D., Hejblum, G., 2009. A systematic review of models for forecasting the number of emergency department visits 26, 395–399.

Wolsey, L.A., Nemhauser, G.L., 1999. Integer and Combinatorial Optimization. John Wiley & Sons, New York.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68, 49–67.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320.

Status of Publication

The following table summarizes the status of publication of the four contributions of this thesis as of 10 April 2020:

Contribution	**Status**	**Co-Authors**
Strategical, Tactical, and Operational Aspects of Bed Bed Planning Problems in Hospital Environments	submission to SSRN planned	–
Combining Clinical Departments and Wards in Maximum-Care Hospitals	published in OR Spectrum on 19.05.2018	Hübner, A. Kuhn, H.
Operational Patient-Bed Assignment Problem in Large Hospital Settings including Overflow and Uncertainty	published in Flexible Services & Manufacturing J. on 10.01.2019	Hübner, A. Kuhn, H. Schäfer, F.
Machine Learning and Pilot Method: Tackling Uncertainty in the Operational Patient-Bed Assignment Problem	submitted to OR Spectrum on 13.02.2020	Grimm, D. Hübner, A. Schäfer, F.

For all three co-authored papers I have engaged in developing the mathematical models, performing calculations, presenting working papers at conferences, writing up manuscripts, and handling review processes. Due to consistency of formatting, spelling, orthography, grammar and nomenclature the versions of the four contributions included in this thesis may slightly deviate from the versions published or submitted to the respective journals. This does not impact content or numerical results.

Declaration of Honour / Ehrenwörtliche Erklärung

I hereby confirm on my honour that I personally prepared the present academic work and carried out myself the activities directly involved with it. I also confirm that I have used no resources other than those declared. All formulations and concepts adopted literally or in their essential content from printed, unprinted or Internet sources have been cited according to the rules for academic work and identified by means of footnotes or other precise indications of source. The academic work has not been submitted to any other examination authority.

Munich, 10 April 2020

Manuel Walther

Hiermit versichere ich, dass ich die schriftliche Dissertationsleistung selbständig und ohne unerlaubte fremde Hilfe angefertigt habe. Ich habe keine anderen als die in der Arbeit angegebenen Schriften und Hilfsmittel benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen kenntlich gemacht. Insbesondere versichere ich, dass ich nicht die Hilfe von Vermittlungs- oder Beratungsdiensten (PromotionsberaterInnen oder andere Personen) in Anspruch genommen habe. Weiterhin versichere ich, dass ich keine früheren Promotionsversuche unternommen, Promotionen abgeschlossen oder die Dissertation in gleicher oder anderer Form in einem anderen Versuch oder in einem anderen Prüfungsverfahren vorgelegt habe.

München, den 10 April 2020

Manuel Walther