Applications of Hospital Bed Optimization

A. J. (Thomas) Schneider and N. M. (Maartje) van de Vrugt

Abstract In this chapter we show typical bed capacity management decisions and how these can be supported using operations research (OR) models. During hospitalization, patients spend most of their time in a bed, situated at a ward. These wards, which include staff, beds, and equipment, are one of the most expensive resources of hospitals. Often patients who stay at a ward receive one or multiple treatments, which usually take place at different departments. Many wards still struggle to accommodate all incoming patients. Without aligned schedules, the flow of patients will fluctuate significantly, and therefore beds at wards will congest. As a result of this "disorganization," staff will experience an unbalanced workload, and wards require more (buffer) capacity to accommodate all patients. With operations research techniques, planning and scheduling of both patient admissions and staff presence at wards can be optimized aiming to reduce variation in the bed occupancy. We also show three case studies using OR in bed management decision-making and discuss success and pitfalls.

A. J. (Thomas) Schneider (✉)
Center for Healthcare Operations Improvement and Research (CHOIR), University of Twente, Enschede, The Netherlands

Department of Quality and Patient Safety, Leiden University Medical Center, Leiden, The Netherlands
e-mail: a.j.schneider@utwente.nl

N. M. (Maartje) van de Vrugt
Center for Healthcare Operations Improvement and Research (CHOIR), University of Twente, Enschede, The Netherlands

Department of Strategy and Innovation, Amsterdam University Medical Center, Amsterdam, The Netherlands

M. E. Zonderland et al. (eds.), *Handbook of Healthcare Logistics*,
International Series in Operations Research & Management Science 302,
https://doi.org/10.1007/978-3-030-60212-3_5

1 Introduction

In this chapter we show typical bed capacity management decisions and how these can be supported using operations research (OR) models. We illustrate practical problems and possible solutions. Furthermore, we show the potential impact of these type of models on capacity decisions in practice and highlight why implementation of the results was successful (see Sect. 4). In our previous work [45] we focused solely on bed occupancy modeling, while in this chapter we point out all capacity management-related decisions for hospital beds.

During hospitalization, patients spend most of their time in a bed, situated at a ward. These wards, which include staff, beds and equipment, are one of the most expensive resources of hospitals and are defined as inpatient care facilities providing care by offering a room, a bed and board [24]. Often patients who stay at a ward receive one or multiple treatments, which usually take place at different departments. As an example, consider admissions at surgical wards, which are strongly determined by the operating room schedule. To optimize ward logistics, the planning and capacity availability should be aligned with other departments' schedules.

Many wards still struggle to accommodate all incoming patients. From our own experiences, we observe that these struggles are foremost a result of unnatural, self-induced variation, caused by unbalanced and unaligned schedules between hospital departments. Without aligned schedules, the flow of patients will fluctuate significantly, and therefore beds at wards will congest. As a result of this "disorganization," staff will experience an unbalanced workload, and wards require more (buffer) capacity to accommodate all patients.

Not all variation is self-induced. Wards typically have unexpected daily fluctuations in the bed census, as a result of changes in patients' health status, staff schedules and/or treatment plans. Admission planners often only take the average length of stay (LOS) into account. Good estimates of patient LOS, when based on multiple patient characteristics such as age and comorbidity, could potentially reduce the difference between the scheduled and realized bed census.

With operations research techniques, planning and scheduling of both patient admissions and staff presence at wards can be optimized aiming to reduce variation in the bed occupancy. When variation can be further minimized, the staff schedules should be adapted accordingly to accommodate for these variations. For example, this may imply that more staff is scheduled to work every Monday or throughout the winter.

Hospital ward logistics typically focuses on two key performance indicators: bed occupancy and blocking probability (e.g., when all beds are occupied). Bed occupancy is an important performance measure, but a universal definition of occupancy does not exist [45]. Next to occupancy and blocking probability, workload is an important performance measure for a ward. Workload also lacks a universal definition but is mainly based on patient (e.g., type of treatment or disease) and staff characteristics (e.g., junior vs. senior staff). Ward resources are scarce, and therefore

hospital management often focuses on maximizing the bed occupancy. When bed occupancy rates are high, the blocking probability increases as well, which implies that patients more often become boarders (e.g., admissions at other wards then dedicated for their medical specialty) or have to be rescheduled. Additionally, it takes significantly more time to find a suitable bed to accommodate a patient. As a consequence, nurses have to treat patients of medical specialties they are not trained for, and doctors' rounds take more time as they have to visit more wards to see their patients. Therefore, striving for high bed occupancy rates at wards competes with both quality of care and job satisfaction.

The remainder of this chapter is organized as follows. We start by explaining the typical ward capacity management decisions in Sect. 2. To support the described capacity decision making, we present the related operations research models in Sect. 3. Next, we discuss case studies where operations research models have made practical impact for ward capacity decisions in Sect. 4. Finally, we discuss future developments and research opportunities in Sect. 5.

2 Ward Capacity Management

The term "planning and control" is most often used for decisions on the acquisition and usage of capacity to efficiently satisfy customer demands [18]. Efficient realization of organizational goals (e.g., satisfied and healthy patients) requires hospital-wide coordination of capacity and flows, by continuously balancing demand and supply. Operations research can give insights to improve the efficiency of capacity and flows, which is increasingly important in these times of rising healthcare expenditures.

To demarcate the scope of capacity management decisions and/or optimization interventions at wards, we use the four-by-four framework of [21] (Fig. 1). The framework hierarchically decomposes managerial levels on one axis, strategic, tactical, and operational (offline and online), and covers different managerial areas at the other axes. An important step in planning and control is setting the length of the scheduling horizon for the different hierarchical levels. At strategic level, decisions are made for at least 1 year but often for multiple years ahead. The operational level is maximally several weeks ahead. Therefore, the tactical level ranges from several weeks to 1 year ahead. On the other axis the framework integrates the managerial planning areas in healthcare: medical, resource capacity, materials, and financial planning. In this chapter we focus on resource capacity planning for both planning and control and operations research models for wards. Following this framework, we use a top-down approach explaining all planning and control decisions at the different hierarchical levels, as higher levels set boundaries for lower levels. Nevertheless, also bottom-up feedback should be in place in practice, so detected deviations and problems can be lifted one hierarchical level upwards for problem-solving.

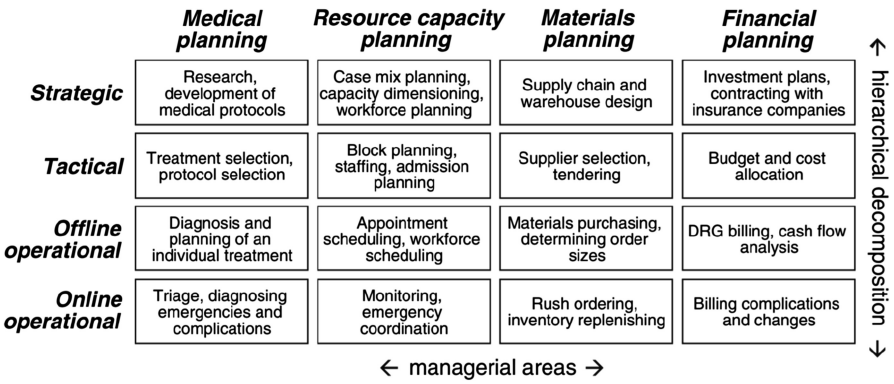

Fig. 1 The healthcare planning and control framework with applications

2.1 Strategic Ward Capacity Management

At the strategic level, the hospital board decides on the hospital's long-term "mission and strategy," the areas on which the hospital aims to focus and excel. Important decisions at the strategic level are the desired case-mix, hospital layout, performance targets, bed capacity, and workforce planning.

2.1.1 The Desired Case-Mix of the Hospital

Based upon the hospital's mission and strategy, the board and/or head of departments determines the case-mix of the hospital; a case-mix is the collection of patient groups a hospital treats. Some hospitals choose a very specific patient group to treat, for example, a breast cancer clinic, while general hospitals have a more diverse case-mix. The preferred case-mix of a hospital determines to a large extent the required capacity.

The case-mix of a hospital can be adjusted by attracting and/or deferring patient groups. As healthcare organizations and their professionals have the duty to care for their patients, a hospital is not allowed to defer a patient group until other regional or national providers agree to treat this patient group. Furthermore, the patient case-mix could also change when a new doctor with a different specialization is hired. The same could happen when specialized doctors leave the hospital.

Adjusting a hospital's case-mix is a complex process as many factors should be taken into account. For example, case-mix decisions could not only affect the required capacity for care delivery but also the education and research possibilities. Another example is that patients not often are treated by a single specialty. So the decision to stop treatment for a specific patient group for a medical specialty could affect the case-mix of many other medical specialties.

2.1.2 Hospital Layout Planning

Based on the mission and strategy, a hospital board decides at the strategic level on the type of wards and rooms that are available in the hospital. One example is the mix between single- and multi-person rooms. Single-person rooms ensure privacy for patients and their family but require more space and result in more walking distance for the staff, and patients monitoring could become more difficult. On the other hand, multi-person rooms are inefficient when patients have infectious diseases and when only same-sex rooms are put in place.

Another aspect of strategic planning for wards is the decision to establish wards for special types of care; acute medical units, intensive care units, cardiac care units, and surgical admission lounges are examples of these dedicated wards. These wards serve a specific patient group based on severity, urgency, treatment or flow (e.g., elective versus acute admissions) and are mainly introduced to improve the quality of care and/or efficiency. As such decisions have high economic impact and take a long time to accomplish, these decisions will affect the hospital logistics for multiple years.

After the global (idea of the) layout of the hospital is determined, all patient groups in the case-mix have to be assigned to wards. An important decision at this level is how many beds are considered and organized as one ward. From a logistical viewpoint, larger wards will result in economies of scale, leading to a higher occupancy with an acceptable blocking probability (see [45]). For this decision, the trade-off between medical and logistical perspectives should be taken into account. Purely from a logistical perspective, if all patient groups could be treated at each bed in the hospital, the bed census at this "single ward system" can be optimally balanced. As a result, nurses in this hospital have to be multiskilled (which is impossible for high-complex care), and doctors will spend more time to visit their patients at different areas of the hospital. From a medical perspective, a more differentiated distribution of patient groups over wards would be optimal, where patient types are clustered according to the skills required for their treatment. A balance between these two perspectives should therefore be found.

2.1.3 Setting Performance Targets

On a strategic level, the hospital board should set performance targets for the hospital. At hospital wards, logistical performance indicators are often the bed census or occupancy, where occupancy can be measured in many different ways [45]. Setting high occupancy targets for wards will, certainly for smaller wards, result in deferring more patients to other wards or hospitals. A different, better performance target would be an upper bound on the percentage of deferred patients and achieving the desired case-mix.

2.1.4 The Number of Beds

When a hospital's desired case-mix is determined, this information is used for strategic planning to forecast the demand for care for the hospital. This forecast is based on aggregated data, trends, and forecasts for the patient population. Using the forecasted demand for care and the set performance targets, a hospital can determine the required capacity to treat these patients. The required capacity is determined on an aggregated scale, such as the total number of required operating theater hours, outpatient clinic hours and ward hours for the upcoming year(s). Based on the aggregated data, the required ward capacity is (re)evaluated yearly.

Typically, the number of physical beds at a ward is higher than the average number of used beds. Each ward should have buffer capacity, to accommodate unexpected peaks in patient arrivals. Often, not all physical beds at a ward are "staffed"; there is no nurse available to treat a patient in that bed. These extra beds ensure, for example, that patients with infection risks can be treated in isolation when the ward has multi-person rooms. Moreover, these beds form a buffer of clean beds if the time between one patient's discharge and another patient's admittance is short.

A ward may also have overcapacity in the number of staffed beds, when the nurse-to-patient ratios do not perfectly match with the expected bed census. The ratio depends on the average "workload" of one patient and denotes the number of patients one nurse can take care of. For example, consider a ward with an expected bed census of 17 beds and the ratios per shift are as follows: day 1:3, afternoon 1:5, and night 1:8. As a result, the shifts requires at least 6, 4, and 3 nurses, respectively. On a strategic level this slack should be taken into account when the expected bed census is translated into the required number of full-time equivalent contracted nurses.

2.1.5 Workforce Planning

At hospital wards often not the physical beds but the number of nurses determines the ward capacity. Once the demand for beds is determined by the case-mix planning, the workforce should be aligned to it. An important capacity management decision that determines how many nurses are required is the nurse-to-bed ratio. For example, an ICU patient has a relatively high workload, and the nurse-to-patient ratio is 1:1, while for a general ward during a night shift the ratio may be 1:16. Workload lacks a universally accepted definition but is generally considered as the relation between the demand (of patients) and the capacity available to fulfill this demand. Workload can be divided into objective (e.g., patient acuity metrics) and subjective (e.g., nurse workload perception) factors [43, 48]. Patient acuity metrics generally consist of activities of daily living, cognitive support,

communication support, emotional support, safety management, patient assessment, injury or wound management, observational needs and medication preparation. Perceived workload is also not universally accepted as measurement for workload but mainly consists of staff characteristics such as age, experience, and educational level. The total workload at a ward is based on the patients' acuity, shift (day, afternoon, or night) and the bed census.

Nurses at a hospital are mostly assigned to one ward or to a few wards that accommodate patients with the same care requirements. At a strategic level, a hospital can decide to flexibly allocate a part of the capacity. For wards, this implies that not all beds are assigned to a certain medical specialty or specific patient group (e.g., organ transplantation patients), but part of the bed capacity will be assigned based on the actual demand for care for each patient group. Flexibility may also imply that a hospital creates a "flex-pool" of nurses; these nurses are often multiskilled and are allocated on a short term (e.g., each morning) to the most busy ward. The advantage of flexible capacity is that a hospital can better adapt to stochastic patient demand. Which part of the capacity is allocated in a flexible way, and on which KPIs the allocation decision is based, is a strategic capacity management decision.

A hospital's desired case-mix also determines the quantity and quality of the required staff to a large extent. However, the translation from case-mix to the number of staff-members is not the same for each hospital. The ever-continuous technological and medical innovations require more specialization from practitioners. In general, more specialization increases the number of involved specialists during diagnosis and treatment, as all specialists are specialized in a small part of a human body and/or specific diseases. Additionally, a hospital's policy with respect to education may change the translation from case-mix to number of required staff-members; junior students usually decrease the capacity due to supervision duties of the staff, while students who are about to graduate can often work without much supervision. Moreover, when the staff-members are involved in research projects, this typically decreases the available capacity for treating patients. Additionally, each hospital has differences in the workforce, which implies that strategic workforce planning should incorporate: influx (training, education, and immigrants) and outflux (retirement or retention) of staff, level of task differentiation (e.g., highly trained versus basic trained staff), and the in- or outsourcing of training and education programs.

Strategic capacity management decisions are always long term and often require major monetary adjustments to accomplish. For economic value and employee satisfaction, yearly changing the ward layout of the hospital is not desirable. However, due to small changes in the case-mix, ward sizes may become inadequate over time. Here, both under- and overcapacity are a problem; see Sect. 4.1. Strategic decisions define the framework at the tactical and operational levels and are therefore to a large extent accountable for the performance of the hospital.

2.2 Tactical Ward Capacity Management

At the tactical level, capacity management decisions focus on organizing the desired case-mix, controlling patient access times and efficient capacity usage via generating master schedules, allocation of flexible capacity, and scheduling bounds which we will explain further in this section. As mentioned earlier, tactical capacity management decisions concern the organization of operations and processes on a "midterm."

2.2.1 Master Schedules

The first step at the tactical level is to divide the total capacity among stakeholders and over the weeks of the year, resulting in a master schedule. Often a master schedule is set for an entire year, but hospitals can gain flexibility to adjust capacity to the patient demand when the scheduling horizon is shorter. For wards a master schedule may result in a weekly schedule where the capacity of each ward changes over time and beds are divided among different patient groups (e.g., elective and emergency admissions). Typically, a master schedule is different for each season. Temporary leaves of staff may require alterations to the master schedule too. Holidays, training, education or internships are examples of temporary leaves and should be planned on the tactical level.

The master schedule of a ward should be aligned with the master schedules of other capacity, such as operating rooms and outpatient clinics, to create a stable flow of patients. Especially for wards that accommodate surgical patients, aligning the master schedules of the operating rooms and wards results in increased efficiency. This alignment is twofold: aligning holiday weeks and balancing bed census. Typically, a hospital has several holiday weeks per year in which both (elective) patients and staff are not available and capacity is reduced for elective (scheduled) care. To prevent a shortage or surplus of beds, the holiday weeks of the operating theater and wards should be aligned. Additionally, by optimizing the operating room master schedule, the postoperative bed census can be balanced better [14]. Balancing the bed census implies that a ward requires less buffer capacity and thus increases efficiency at a ward and reduces the risk of cancelling a surgery due to a lack of postoperative beds.

2.2.2 Flexible Allocation of Capacity

Patient demand is usually fluctuating; thus, it may be beneficial to adjust part of the capacity throughout the year. For the total flow in the hospital, admitting a similar number and type of patients each week is optimal. However, due to staff holidays and stochasticity in patient arrivals, this is a difficult aim. For each patient group (or aggregated for each medical specialty), it is therefore beneficial to periodically

evaluate the available capacity and make minor adjustments where necessary and possible. For wards, this could imply asking nurses to work on a different ward for several weeks or doctors to, for example, help at the ward instead of working at the outpatient clinic.

Hospital management can decide on the strategic level to partly reserve capacity and allocate this on a regular basis, e.g., monthly. At the tactical level, this capacity may be allocated for several weeks in advance. When applying flexible capacity allocation, it is important to have consensus among all stakeholders about the parameters and performance indicators upon which the allocation will be based and on the scheduling horizon on which the flexible capacity can be allocated. For wards, flexible capacity could imply that nurses from the flex-pool are assigned to a specific ward. Moreover, downstream resources should be taken into account when allocating staff from a flex-pool to align patient loads. For example, allocating flexible operating room time affects the bed census at the postoperative wards; thus, these decisions may require additional nurses at the postoperative wards.

At most hospitals, staff rosters are generated several months in advance, and therefore staff planning is performed on tactical level too. These rosters only state which shifts and days an employee should work and do not specify the department, bed, and/or patients. The latter, detailed scheduling, is performed on the operational level. This provides the departments additional flexibility in staff allocation.

2.2.3 Regulating the Demand for Care

In order to balance patient flow and optimize efficiency, at the tactical level hospital management can decide to formulate rules for patient scheduling. Such rules may, for example, state the minimum and maximum number of elective patients that may be admitted to a ward per day. Another example is stating a maximum on the number of surgeries scheduled at the same day that require an ICU bed. These rules can be ward specific, medical specialty specific, or may hold for the entire hospital.

Tactical capacity management decisions are crucial to efficiently organize patient care and flows, especially at the interface between different types of resources in a hospital. From our own experience, this level is still underdeveloped in many hospitals.

2.3 Operational Ward Capacity Management

The operational level is divided in offline (service at a later point in time) and online (instant service) capacity management decisions. Compared to the other two levels, the operational level has very limited possibilities to adjust the capacity to the patient demand. The online level comprises the actual patient (room and bed) to staff scheduling and ad hoc decisions, such as replacing ill staff-members and admissions of emergency patients.

2.3.1 Patient Scheduling

Patient scheduling on the operational level comprises deciding each elective patient's admission date and ward. It should be taken into account that each admittance and discharge imply a workload peak for the nurses. Additionally, it is important to take the expected urgent patient admissions into account, as scheduling too many elective patients results in deference of emergency patients. Moreover, a patient schedule should minimize the number of in-hospital patient transfers, as each transfer could be a risk for the quality of care. Therefore, a patient schedule should, for example, take into account that some wards close beds during the weekends and therefore do not accept admissions that are expected to stay longer than Friday as the ward is closed during the weekends. Yet, when patients need to stay after Friday, they have to be transferred to other wards. Accurate predictions of the length of stay (LOS) are therefore crucial. Some hospitals/wards adjust the patient schedule one week in advance, based on the actual bed census and LOS predictions of the currently admitted patients.

At the online level, a ward manager may decide to transfer a patient with relatively good health to a ward with a lower care level or to another hospital, to reserve capacity for high-care patients. In practice, patients are typically admitted to their medically preferred ward when there is available capacity, and ward managers start to transfer patients to "second-best" alternative (also called "overflow") wards when capacity runs out. As a consequence, patients may wait for a long time at wards that are not medically preferred before a bed is available, as another patient needs to be transferred first. To decrease the time until a patient is assigned to a bed, ward managers may already transfer patients when there is still available capacity, to reserve enough capacity for new patients. By focusing on the expected discharge date at the moment of arrival, the LOS will decrease (this is also called discharge management). The decision on which patients should be transferred is often difficult, and optimizing this decision-making process may improve patient waiting time, quality of care, and even hospital revenues significantly. Hospitals may also apply admission control to make sure enough capacity is available for the patients that need it the most and/or benefit from it the most, especially when there are multiple hospitals in the proximity.

2.3.2 Staff Scheduling

At the operational offline level, staff is assigned to a specific ward several weeks in advance. When the hospital has a flex-pool of nurses, these nurses may be allocated to specific wards at the operational level. Some hospitals allocate these nurses several weeks in advance, based on long-term illness of staff, short-term staff leaves or forecasted patient demand. Hospitals may also decide to assign nurses from the flex-pool in an online way, which implies that a nurse is assigned to a ward at the beginning of a week or even each shift.

At the online level, nurses are assigned to patients. This scheduling task is performed before each shift starts. Next to the number of patients present, also the "type" of patient is important to optimize the nurse-patient assignment. As patient acuities and staff characteristics vary over time, nurse-patient assignments should be optimized by distributing the workload among available nurses on the operational level.

Although there is little room to adjust capacity to actual demand on the operational level, we have shown many capacity management decisions on this level that can further optimize patient care delivery. As improvement on this level requires relatively small adjustments in terms of work routine and/or investments, they are relatively easily to implement. Therefore, both management and staff can fulfill the potential of these improvements at any moment.

2.4 Feedback Between the Hierarchical Levels

As common in literature, we have used a top-down approach for discussing all hierarchical control levels at wards. As mentioned earlier, healthcare processes and planning deal with stochasticity, and therefore unforeseen situations often occur. Monitoring systems should be in place to detect deviations from scheduled care processes. Using data from electronic health records, software can easily detect, present, and even predict these deviations. It is important to note that some data has to be put in manually (e.g., the expected discharge date) in order to accurately detect deviations. When a deviation is detected or predicted, planners and ward management can proactively arrange adjustments in capacity and/or demand.

When detected deviations cannot be solved within the managerial boundaries of the level where the unforeseen situation occurred, the deviations should be escalated. Bottom-up feedback loops provide escalation channels to lift problem-solving to higher hierarchical levels. It must be clear for each level when detected deviations have to be escalated. An example for escalation could be regularly occurring peaks in postoperative elective patient arrivals; the master schedule of the operating theater should then be reconsidered in order to balance the postoperative arrivals at wards. In general, recurring problems may require structural redesign of processes and thus require decision-making on a higher hierarchical level. Therefore, escalation channels are an important component of the planning and control cycle for resource capacity planning.

3 Operations Research Models for Wards

In this section we reflect on OR models that can be used for analyzing ward capacity management decisions. We follow the same hierarchical approach as the previous

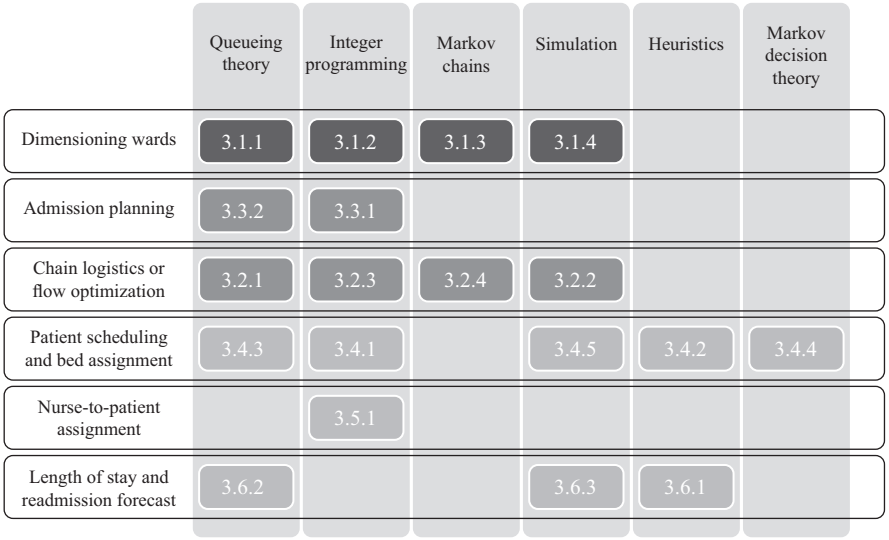

Fig. 2 Overview of Sect. 3

section and show the capacity management decisions covered in this chapter and used OR techniques from literature to analyze these types of decisions in Fig. 2.

One important optimization application at wards is nurse staffing. Although the physical capacity of a ward is determined by the number of beds that are present at the ward, in most hospitals the number of nurses present at the ward determines to a large extent the number of patients that can be accommodated. Many departments schedule the same number of nurses each shift or marginally adapt the nurse schedule to the bed demand. The topic is elaborated upon in chapter "Bed Census Predictions and Nurse Staffing" and is therefore not discussed in this chapter.

3.1 Dimensioning Wards

Finding the optimal capacity of a ward by allocating patient groups among wards is a typical strategic decision. In the literature, dimensioning decisions are based upon queueing models, Markov chains, simulation, goal programming, and mixed integer programming models. Below we will evaluate these approaches and provide some examples from the literature.

3.1.1 Queueing Theory

The Erlang loss and infinite server queueing models are by far the most-used models to determine the best dimension of hospital wards. With easy-to-use tools available, such as the Queuing Network Analyzer (see [56]), hospital practitioners are able to analyze decisions with queueing models. The examples provided by [45] and in the case study presented in this chapter in Sect. 4.2 demonstrate the value of these basic models for dimensioning hospital wards. Another advantage of the Erlang loss queue and infinite server queue is that these models are insensitive to the distribution of the length of stay; obtaining an average LOS from hospital data is enough for the analysis. Sophisticated data analysis to generate input data is therefore not required for these models. The basic queueing models do not encompass all hospital ward dynamics. For example, they do not encompass nonhomogeneous arrival and discharge rates, while in reality scheduled patients only arrive and discharge during the day. Another example of misrepresentation is that in practice patients are often not "blocked and lost" if all beds at their medically preferred ward are occupied upon their arrival, which implies that queueing models underestimate the bed occupancy. Gallivan and Utley [15] demonstrate that for an infinite server queue with piecewise stationary Poisson arrivals, the resulting model is easy to analyze. However, most queueing models become intractable with time-varying arrival and/or service rates. Additionally, feedback and overflow are typically difficult to analyze, as shown by, for example, [46], for a small network of an operating theater and an intensive care unit (ICU). To increase the predictive value of the model, Williams et al. [53] consider an Erlang loss queue in which the arrival rate depends on the number of occupied beds, to reflect that less patients are admitted to the ward when it is almost full. Bekker and de Bruin [3] analyze an infinite server queue with time-dependent arrival rate and use the square-root staffing rule to dimension an ICU.

3.1.2 Integer Programming

Queueing models alone require a trial-and-error approach to find optimal capacity. To overcome this problem, queueing models can be incorporated into a mixed integer programming approach. van Essen et al. [47] analyze three approaches assigning patient group clusters to wards. The exact approach uses the Erlang loss model to determine bed capacity given a blocking probability and an ILP is used to determine which patient groups should be clustered and assigned to a wards. The second approach uses an approximation of the Erlang loss model by a linear function for the required number of beds followed by the ILP for the clustering process. The last approach uses the exact formulation of the Erlang loss model for the number of required beds and a local search heuristic forming the clusters. Another example of combining queueing models with optimization models is given by [38], who use similar approaches as in [47] to determine the bed capacity for a network of maternity clinics. Pehlivan et al. [38] also linearize the blocking probability and bed

census of the Erlang loss model and also analyze interactions between clinics with a mathematical model. The queueing model formulas can also be incorporated in a goal programming approach. Li et al. [32] use this approach to allocate a number of beds at each ward, to ultimately optimize multiple objectives set by the hospital management.

Oddoye et al. [37] use simulation to relate the capacity (beds, nurses, and doctors) of a medical assessment unit to queue lengths for patients and incorporate this into a goal programming model.

3.1.3 Markov Chains

Predicting the bed census using Markov chains may result in higher accuracy compared to a queueing approach, as time-varying arrival and discharge rates may be incorporated in such models. Kortbeek et al. [30] invoke a Markov chain to predict the hourly bed census, which includes postoperative surgical patients, emergency admissions, and overflow patients to and from other wards. Using the steady-state distribution, the authors obtain an expression for the 95% percentile of the bed census.

Markov chain models are also applicable in transient analyses; for example, Broyles et al. [8] predict the ICU bed census invoking a transient Markov chain analysis with maximum likelihood regression.

Using Markov chains almost every desired detail can be modeled. However, including more details into the models quickly makes a Markov chain intractable.

3.1.4 Simulation

With simulation, all features of hospital wards imaginable can be incorporated, which makes this type of modeling sometimes the best or only option to model a ward. Holm et al. [23] use a simulation model to relate the capacity of a ward to the bed census for several wards for all possible numbers of beds and heuristically assign beds to the different wards using these relations. The model is evaluated using data of a university medical center. VanBerkel and Blake [49] analyze multiple scenarios to solve waiting list issues, of which redistributing beds among the wards is one.

Transient analyses are also possible while using simulation models: Zhecheng [55] presents a simulation model to obtain short-term predictions based on the specific characteristics of the current patient population present at the ward.

Simulation models require labor-intensive data analyses to generate input parameters, developing time, often require complete enumeration and output analyses. Furthermore, strategic analyses do not always require all details, and therefore simulation models are not an obvious first choice to determine the best capacity of a ward. Both academics and professionals should therefore keep in mind the trade-off between required level of detail and the required amount of building and running

time and the value of the outcomes when using simulation techniques to analyze capacity management decisions.

3.2 Chain Logistics or Flow Optimization

So far we have highlighted research that mainly focuses on a particular step in the patient care pathway: clinical treatment at inpatient wards. However, the inflow of inpatients is often determined by other hospital departments. Especially for wards that accommodate many surgical patients, the operating theater schedule determines to a large extent the bed census at the ward. For wards that accommodate many urgent patients, the emergency department and acute admission unit influence the bed census. Vanberkel et al. [50] provide a survey of healthcare models that encompass multiple departments. Interestingly, achieving optimal logistical flow through a hospital may result in suboptimal use of capacity of individual resources. In this section we will highlight literature on queueing, simulation, MIP, and Markov chain models that mimic the interaction between multiple departments.

3.2.1 Queueing Theory

Queueing networks are useful in relating different capacity levels to certain performance measures such as blocking probability. For example, Zonderland et al. [57] analyze multiple scenarios for a network of an emergency department, acute admission unit, and two wards, in which the acute admission unit may function as an overflow for the other three departments. They observe that with the used setting the arrivals of urgent patients can be increased at the expense of decreasing elective arrivals (the increased influx of emergency patients is higher than the decreased number of elective arrivals).

Litvak et al. [33] tackle the problem of deferred acute intensive care patients as a result of capacity problems. They show that regional cooperation between multiple ICUs results in higher acceptance levels for these patients. The authors approximate the blocking probabilities in an overflow network with the equivalent random method and the Erlang loss queue. Setting a threshold to this blocking probability, they determine how many beds each ICU in a certain region should reserve (make available for regional patients) such that all acute intensive care patients in the region can be accommodated promptly.

3.2.2 Simulation

Simulation models are also used to analyze patient flows through multiple hospital departments and to determine the effect of changes in, for example, the capacity. Optimizing patient flows at individual departments of a hospital may lead to

disturbed patient flows at other departments, as the bottleneck in the patient flow may shift to another department [29].

Schneider et al. [42] use simulation to analyze the flow of emergency patients among the three departments: ED, acute medical unit, and inpatient wards. The hospital of their case study has great difficulties accommodating all emergency admissions. Using heuristics they optimize the number of allocated beds per inpatient ward for emergency patients that need to stay longer than is intended at the acute medical unit.

Another example of a simulation study in this area is Mustafee et al. [36], who investigate different strategies preventing patients to occupy high-care beds unnecessarily long due to the unavailability of beds with a lower level of care. Day et al. [12] investigate the effect of adding capacity and a different discharge policy on the patient flow at a pediatric surgical center.

3.2.3 Mixed Integer Programming

Mixed integer programming models are often developed to optimize the master surgery schedule (MSS) of the operating theater. Fügener et al. [14] optimize the MSS while minimizing the probability that overcapacity is necessary to accommodate all patients at the postoperative wards. Based on which surgical specialty is assigned to which time slot in the MSS, they analytically express the bed census distribution function for each ward.

An operational offline approach is taken by Gartner and Kolisch [17], who invoke an MIP model to determine the admission dates for patients that require care at multiple departments.

3.2.4 Markov Chains

A Markov chain approach was invoked by Isken et al. [27] to model multiple patient pathways at an obstetrics department with multiple wards. The obtained expressions are incorporated into an MIP model to optimize the schedule of elective patients.

3.3 Admission Planning

After the capacity dimensions are set for wards, demand and capacity can be optimized on a midterm horizon through admission planning and nurse rostering, respectively. As mentioned in the introduction, admission planning generates a blueprint schedule that schedules the different patient groups and not individual patients and/or treatments. For this type of optimization, mixed integer programming (MIP) models are mostly preferred in the literature.

3.3.1 Mixed Integer Programming

In our previous work, we used a mixed integer programming approach to develop a tactical schedule for a weekday ward (see [45]). Weekday wards only admit elective patients during weekdays and close during the weekends. All patient care is delivered according to strict protocols, which results in highly accurate treatment time and LOS predictions. Typically, treatments with a longer LOS are scheduled at the beginning of a week and shorter treatments later during the week, to ensure that all patients are discharged before the weekend.

Helm and Van Oyen [22] develop two infinite-server queueing models (one for emergency arrivals and one for elective arrivals) to determine the bed census that results from any admission plan for regular wards. Based on these bed census, an MIP model minimizes the blocking probability of emergency arrivals, the cancellation probability of elective arrivals, and the average number of boarders (patients who have to wait for their preferred ward) in this tactical admission plan.

Bekker and Koeleman [4] use a quadratic program to obtain daily quota for the number of admissions to a ward to minimize the variability in the bed census (e.g., quota planning). In the quadratic program the bed census is modeled using a GI/G/∞ queue with a heavy traffic approximation, and the authors present an approximation for the bed census of a ward that experiences a non-Poisson arrival process. The aim is to generate rules of thumb for management and planners based on the results of their model. They conclude that quota planning has the most impact on the bed census variability (e.g., smooth bed census during the planning horizon). Using quota planning the arrival process of admissions at wards will be more stable. A stable arrival process results in a more stable bed census compared to high variable arrival rates. The next rule of thumb is to schedule arrivals given the number of available (or closed beds) during the planning horizon. Given the absence of admissions during weekends, beds can be closed, or patients with a longer LOS can be scheduled on Friday to improve the bed census during weekends.

Typically, patients require other types of resources during hospitalization, such as the operating theater or diagnostic facilities. Therefore, the patient schedules at these resources affect each other. Incorporating many different capacity types and patients following uncertain treatment paths (see Hulshof et al. [25]) invoke an MIP approach to optimize the number of admitted patients per time period. The tractability of the MIP appears insufficient to optimize realistic instances, and therefore the authors turn to approximate dynamic programming [26].

3.3.2 Queueing Theory

Queueing approaches may be used to determine the best number of beds that should be reserved for a certain patient type. For example, Litvak et al. [33] investigate a network of ICUs that all reserve some capacity to admit emergency patients in the region using the equivalent random method. As ICU capacity is scarce and costly, it is typically utilized maximally, which results in blocked emergency patients and

cancellations of scheduled patients. The analysis shows that when multiple regional ICUs cooperate as a network, they can increase the acceptance level of emergency patients with a smaller total number of beds compared to the setting of individual ICUs. Mandelbaum et al. [35] present another application of queueing models where they balance the bed census of wards with a similar level of care by considering routing algorithms for patients from the emergency department to wards.

3.4 Patient Scheduling and Bed Assignment

The tactical admission planning results mainly in a blueprint schedule for patient admissions at wards. In the next planning phase (e.g., operational planning), actual patients are scheduled and assigned to available beds. The tactical blueprint serves as a guideline for scheduling patients. In some circumstances (e.g., patients that have been scheduled and/or availability of staff), management and planners can deviate from this blueprint. This would not be preferable as the blue print of downstream resources should also deviate. Using optimization, patient admission dates and bed assignments can be chosen such that the number of beds that is required to treat all patients is minimized, or the variation in bed usage is minimized. Additionally or alternatively, the number of patients that receive treatment within their preferred access time window can be maximized.

Optimizing bed assignments will have the largest impact when the medically preferred ward has multi-person rooms or when there are several wards with adequate level of care. For multi-person rooms, for example, patients with infectious diseases and "same sex in one room," rules may complicate the room assignments. Basically there are two decisions in this type of problem: (1) shifting admission dates and (2) transferring patients between wards according to medical preferences.

Below, we highlight literature with respect to patient admission scheduling, bed assignment, and admission control. Here we exclude literature on (surgical) patient scheduling; for more insight on surgical scheduling, the reader is referred to [58]. Sets of benchmark instances for the offline optimization of bed assignments[1] and the patient admission scheduling problem[2] are available online.

3.4.1 Mixed Integer Programming

MIP models are incorporated in an online decision support system to optimize bed assignments. For example, Schmidt et al. [41] and Vancroonenburg et al. [52] determine the optimal ward and/or bed assignment for each patient with respect to the bed census among all wards, the adequate level of care for as many patients

[1] https://people.cs.kuleuven.be/~wim.vancroonenburg/pas/

[2] http://satt.diegm.uniud.it/index.php?page=pasu

as possible, and the number of transfers that are required during treatment. Ben Bachouch et al. [5] also apply an MIP approach to assign patients to beds, where elective patients request a time window in which they require treatment, whereas for emergency patients this window starts at the current time and is equal to the length of stay.

Bed assignment decisions may also be optimized in an offline setting, where all patients scheduled for admission are assigned to beds in an optimal fashion. Braaksma et al. [6], for example, use an MIP to solve an operational offline patient scheduling and bed assignment problem at a weekday ward, in which they optimize over all medically preferred patient access times. Guido et al. [20] present a heuristic based on an MIP model to satisfy as many bed assignment constraints as possible in an offline optimization model while taking into account that some patients may require care from multiple medical specialties.

3.4.2 Heuristics

The patient admission scheduling problem including all constraints on bed assignments and patient access times is proven to be NP-hard by [51]. Therefore, recent literature is more frequently applying heuristics to improve the admission schedule. Kifah and Abdullah [28], for example, apply a "great deluge" algorithm to optimize admission dates and bed assignments in an offline setting. In this paper, the used neighborhood-search algorithm is compared to other heuristics known in the literature and concludes that this great deluge algorithm can compete with more familiar heuristics such as simulated annealing.

3.4.3 Queueing Theory

Griffiths et al. [19] investigate an operational admission control for an ICU using an Erlang loss queue with both elective and emergency arrivals. For analyzing the bed census, they show that both the arrival streams and service rates can be combined into a single queue with multiple servers (e.g., an $M/G/c/c$ queue). The authors analyze the system by controlling the elective patient admission dates based on the bed census using Euler's method to analyze the loss queue with time-dependent arrival rates. Using historical data, they show that it is possible to estimate the most probable level of bed occupancy for several days in advance, given the bed occupancy on the current day. In addition, the model is able to predict the expected split between emergency and elective patients over the forthcoming days. Based on the expected bed occupancy in the near future, staffing levels can be adjusted.

3.4.4 Markov Decision Theory

Optimizing patient scheduling decisions using a Markov decision approach typically results in complicated scheduling policies that are difficult to implement in practice. For example, Barz and Rajaram [2] model patients with stochastic length of stay at multiple hospital resources (e.g., beds and operating rooms) such that emergency patients can always be admitted and elective patients are delayed or deferred. Even the approximate dynamic programming model was not solvable within the set time limits for realistically sized instances, and the authors evaluate some heuristics based upon the results for small instances.

A Markov decision process approach is also used by Yang et al. [54] to decide which surgeries have to be rescheduled such that the ICU capacity is not exceeded. The authors base a heuristic solution approach upon the obtained optimal policy and apply it to cardiothoracic ICU data of surgery requests. It appears that the heuristic policy outperforms the current admission policy significantly.

Markov decision models are also used to determine an optimal bed assignment policy in an online setting. For example, Thompson et al. [44] consider a hospital in which patients should be admitted to a bed at their medically preferred ward or one of the predetermined alternative wards. Dai and Shi [11] consider a similar problem and use approximate dynamic programming to optimize the decision whether to assign patients to their medically preferred ward or to the "second-best" ward.

Transferring patients during their stay may optimize the bed assignments and shorten the time between admittance and bed assignment. These transfer decisions are often optimized together with the assignment of newly admitted patients. Transferring patients during their stay could be optimal from a bed census perspective. One could ask if other factors (e.g., quality of care, patient condition, and staff workload) should also be considered implementing such decision rules.

3.4.5 Simulation

Simulation models are used to investigate a number of bed assignment policies for specific hospital case studies. For example, Braaksma et al. [7] study different policies to reserve beds for patients that are about to be brought to the operating theater. Landa et al. [31] evaluate different policies of reserving beds for patients that are admitted to the hospital through the emergency department.

In Sect. 4.1 we describe a simulation model to investigate how many beds should be reserved for high-care patients, which implies that patients with lower care requirements should be admitted to a ward that is not their medically preferred one.

3.5 Nurse-to-Patient Ratio

The physical beds at a ward are often not the limiting factor in the number of patients that can be accommodated. The number and type of nurses and the specific patients present at the ward determine whether there is capacity for new admissions. The nurse-to-patient ratio says how many "average" patients one nurse can take care of; if there are 5 patients with a high care demand, a ward can be full, while with 15 patients with a low care demand, there may still be available capacity. An acceptable workload is important for the well-being of nurses and the quality of care.

3.5.1 Integer Programming

To balance the workload fairly among the nurses, mostly linear programming approaches exist in the literature. For example, Acar and Butt [1] consider patient acuity scores and travel distances for the nurses in optimizing the nurse-patient assignments. Braaksma et al. [6] additionally consider the continuity of care, education, and patient or nurse preferences in the optimization. Pesant [39] applies a goal programming to optimize the nurse-patient assignments, extending an MIP approach from [43]. In this model, patients have a nurse-dependent acuity, motivated by differences in experience, training, or preferences of the nurses.

3.6 Length of Stay and Readmission Forecast

The length of stay of patients in a hospital is typically not known exactly before the patient is admitted and sometimes even not known exactly the day before the patient may be discharged. Moreover, when a patient is discharged, there is always a possibility that the patient has to be readmitted for further treatment. The time a patient is medically ready to be discharged and the readmission probability are useful in the patient scheduling process, as these determine how many new patients may be admitted. In recent literature, we see queueing theory, simulation, machine learning, and regression approaches, of which we provide examples below.

3.6.1 Heuristics

An example of a machine learning approach (random forest model) is used to forecast the length of stay of obstetric patients, using information on a patient's medical history from the electronic medical records [16]. Roumani et al. [40] forecast the readmission probability for a cardiac ICU, for which they compare a support vector machine, decision trees, and a logistic regression approach. The results of these studies may be implemented in a decision support tool and provide

guidelines to practitioners which clinical measurements that indicate a relatively high risk of prolonged length of stay or an increased readmission probability.

3.6.2 Queueing Theory

Queueing theory is, for example, used to investigate the effects of different discharge policies at an ICU [34]. These authors investigate the practical implications of the best policies using simulation. When a patient needs to be admitted to the ICU at a moment all beds are occupied, typically the "most healthy" patient is discharged to a ward with a lower level of care; optimizing such decisions may improve the quality of care significantly.

For a general ward, Chan et al. [9] develop an infinite server queue in which a server may only be released after an inspection, which mimics the final doctor visit before a patient may be discharged. The results of the queueing analysis indicate that inspections should be at equidistant times and additional inspections have a decreasing marginal reward.

3.6.3 Simulation

An example of a discrete event simulation model is given by Crawford et al. [10], who analyze the effect of different discharge strategies on the readmission rate and emergency department crowding for a complete hospital. The authors conclude that a more "aggressive" discharge policy that discharges patients as early as possible increases the readmission rate significantly.

3.7 Conclusion

We have shown a broad overview of situations where patients and staff at hospital wards can benefit from operations research analyses. Obviously, each different research question may require a different modeling approach, but, as we have demonstrated above, many models are applicable to analyze similar capacity decisions. Queueing models are fast to obtain estimates and are therefore applied often as a first indicator of, for example, the required capacity. Typically, Markov chain models are less "general" compared to queueing models, as they are more difficult to reapply to other wards but are easier to model transient behavior and ward-specific patient admissions, discharges, and transfers. Similar to Markov chain analyses, mixed integer programming approaches are relatively case study specific. However, using MIP approaches processes and schedules may be optimized, although a complicating factor often is the stochasticity in healthcare processes. Machine learning and regression approaches are very useful to analyze large amounts of hospital data and are increasingly used to assist medical decision-making at wards.

4 Impact in Practice of OR at Wards

In this section we present three case studies that we conducted at our partnering hospitals. In all projects the hospital has implemented the results of the research. In these short case studies we focus on the practical approach that was taken and the insights from implementation. Each case study gives unique insights on success factors, pitfalls, and lessons learned.

4.1 Case Study I: Balancing Bed Census

Two medical wards at the Jeroen Bosch Hospital (JBH) experienced unbalanced bed occupancies during 2012 and the first months of 2013.[3] At the neurology department of the JBH, patients' LOS was reduced significantly, resulting in more slack capacity. At the same time, the department of internal medicine experienced increasing numbers of patients, resulting in crowded wards and many patients being deferred to other wards.

Both over- and under-capacity are a problem for wards. In case a ward has under-capacity, patients cannot always be accommodated at the medically preferred ward. As a consequence, patients of one medical specialty are placed at many different wards, and doctors spend much time visiting all their patients. Having overcapacity is a problem for hospital staff as many patients from other medical specialties are likely to be placed at the ward. As a consequence, nurses from the ward have to care for patients for which they were not fully trained and may experience a high workload if they feel incompetent to treat patients from other medical specialties. In both scenarios, patients do not always receive the best possible care, which increased the willingness of all stakeholders for solving this problem.

4.1.1 Project Organization

In accordance with the list of factors in [45], at the start of this project we commissioned a steering group consisting of all stakeholders in this problem: a neurologist, a specialist internal medicine, an administrator from the patient admission scheduling office, and all involved ward managers. The hospital management made this steering group responsible for finding a solution for the over- and under-capacity at the wards and made one organizational and a healthcare logistics advisor member of the steering group. One representative of the highest management level below the board of directors was made chair of the steering group. The neurologist and internist were selected based upon the trust and goodwill they had from their peers.

[3]This case study was conducted by the author, among others, N.M.(Maartje) van de Vrugt in the role of healthcare logistics advisor.

These representatives were not necessarily the heads of department, since it was required that these doctors spent time on the wards and experienced the problems on a daily basis.

The first meeting of the steering group started with getting to know all members of the group; although all stakeholders work on closely related topics, typically they do not often meet and/or talk to each other. The group discussed to what extent they experience a problem at the ward or during patient scheduling. Supporting this discussion, the logistical advisor presented the results of a data analysis with information on (1) the bed occupancy of all hospital wards, (2) the bed requirement per medical specialty, and (3) the number of patients per medical specialty that was not treated at their medically preferred ward. The (fully anonymized) data that was used for this analysis was routinely collected hospital data on admittance and discharge date, medical specialty, and ward. The data analyses objectified the discussion significantly. For example, at the neurology ward the nurses experienced a high workload, and the hospital data confirmed that the nurses had to take care of relatively many patients from other medical specialties, which increased the experienced workload.

4.1.2 Analysis of Possible Interventions

The result of the first session was that the steering group wanted to investigate two possible interventions:

1. Opening an acute medical unit (AMU), and
2. Reassigning medical specialties to wards,

Invoking an $M/G/s/s$ queue, the required bed capacity to achieve at most 5% blocking probability was determined for each specialty. This analysis confirmed the belief of the steering group that the distribution of beds among the specialties was not adequate but adding capacity to the total system was not necessary. For intervention 2, each of the possible scenarios required serious rebuilding of units or medical specialties being split up among multiple wards. Rebuilding several wards would be costly and would take several months. Therefore, the steering group decided to discard this intervention.

The effects of intervention 1 were analyzed for several scenarios using an $M(t)/M(t)/s/s$ queue [45]. The conclusion of this analysis was that the acute medical unit would not be beneficial for the hospital's case-mix, and the steering group discarded this intervention.

At this point in the project, the steering group was looking for new possible interventions and decided to investigate the possibility of creating an overflow ward for internal medicine at the neurology ward. In the analysis of intervention 1, each doctor had to determine a list of diagnoses for their specialty that were eligible to be treated at an acute medical unit. This list consisted of diagnoses that required a relatively low care level, and each acute patient with a diagnoses from the list would

be admitted to the acute medical unit. With minor moderations to the list by the internist, the list was adequate for the eligible overflow patients.

Since the admittance data was anonymized, an exact analysis of the overflow ward was not possible. Financial hospital data revealed which part of all internal medicine patients had diagnoses from the list, which was used to estimate the total overflow bed requirements. This number of beds was high enough to alleviate the pressure on the internal medicine ward and was low enough to be accommodated at the neurology ward.

4.1.3 Choosing an Intervention

Based on this promising result, the organizational advisor helped doctors and nurses from internal medicine and neurology to investigate what would be required, for example, in terms of skills, education, and doctors' rounds at the wards. The most important decision at this level was how often the internists would visit the overflow patients, which medical decisions were allowed to be made by neurologists, and when an internist should be called for assistance. Based on these discussions, nurses and doctors were confident that the quality of the provided care would be good for the overflow patients.

Additionally, the logistical advisor conducted a simulation study in which historical data was used to determine the best policy to start and stop overflowing patients. In this simulation, each patient was randomly eligible for the overflow ward. The steering group requested this additional research as the neurology ward manager feared that, due to the overflow patients, not enough beds would be available for neurology patients. Several overflow policies and their effects were presented to the steering group.

Based on all gathered results, the steering group decided to implement the overflow ward, using the policy: only overflow if both (i) three or fewer beds are available at the internal medicine ward and (ii) two or more beds are available at the neurology ward. In September 2013 the intervention was implemented at the hospital.

4.1.4 After Intervention

In January 2014 data analysis and interviews with the staff showed that the intervention had the desired effect: the neurology ward accommodated more internal medicine patients (on average 2.5 beds) and less patients of other specialties. Both effects were statistically significant. Additionally, the internists reported a reduction of the time required for their rounds, and the neurology nurses experienced a reduction of the fluctuations in the workload and were confident to deliver high quality of care. A downside of the implementation appeared to be a higher workload at the internal medicine ward, as many of the "easier" patients were admitted to the neurology ward.

4.1.5 Lessons Learned

The success of this case study relates to the fact that the interventions were proposed by clinical leaders and on objectifying the effects of these interventions prospectively using data. Based on these analyses, the steering group was able to choose the most promising intervention to implement. The higher management let the steering group choose what interventions to investigate but had set a clear target to find a solution for the problem. This autonomy was greatly appreciated by the steering group.

A very important part of the project was checking all assumptions and data analyses with nurses and doctors working at the wards. The goal of many data validation discussions was to come to an agreement that the data indeed reflected what happened in reality at the wards. A lengthy discussion about data or assumptions during steering group meetings would be undesirable, as this would lead the group away from finding a solution.

During the project there was an emphasis on finding a solution that all steering group members and involved staff would consider a clear improvement of the current situation. To this end, next to data analysis, a thorough risk analysis was done for every intervention the steering group suggested. It was important not to ignore any of the concerns of the steering group, as this would decrease the willingness to cooperate in implementing the intervention. For each of the concerns raised, if possible data analysis was performed, and the steering group took time to discuss all concerns thoroughly, until either the issue was alleviated or the corresponding intervention was discarded. One example is that the neurologists were concerned that the internal medicine patients would displace the neurology patients. This issue was alleviated by a simulation analysis with multiple scenarios, which eventually lead to a decision rule for the patient admission planners.

Before the project started, higher management had emphasized with the steering group members that this project was initiated to improve both quality of care and employee satisfaction. When discussions within the steering group were boiling down to competing interests of the individuals in the steering group, the chair of the meeting reminded everyone to stay focused on the quality of care and employee satisfaction. In all discussions this reminder sufficed to find a common goal and, eventually, a solution to the problem. The autonomy of the steering group and the iterative process of testing possible interventions resulted in an intervention supported by all involved staff. This support was the key to the success of the intervention. Moreover, the intervention proved to be effective in reality, which was the ultimate goal of the project.

4.2 Case Study II: Dimensioning Wards

The Leiden University Medical Center (LUMC) dealt with multiple logistical problems at its wards.[4] These problems related to the small wards in terms of number of beds and a medically illogical distribution of patient groups among wards. This resulted in rising numbers of refusals at the emergency department and increasing waiting lists. Small wards have more difficulties coping with variability such as arrivals and LOS as it has relatively more impact. Therefore, small wards will often have over- and under-capacity.

4.2.1 Project Organization

In 2014 the hospital board of directors decided to redistribute patient groups among wards and re-dimension wards. Therefore, a project was initiated with a steering group consisting of a management director (project lead), a project manager, all care managers, an organizational consultant, a change management consultant, a human resource consultant, and a healthcare logistics consultant. Furthermore, there were multiple topics for further analysis defined and one working group for each topic. To overcome the earlier mentioned logistical problems, the project had to implement the following interventions:

- Redistribute patient groups among wards for long stay patients (e.g., a LOS of at least 5 days).
- Introduce a ward for short stay patients (e.g., LOS of less than 5 days).
- Introduce a ward for day treatments (e.g., LOS of at most 8 h).
- Introduce a ward for acute admissions (e.g., an AMU).
- Merge ward staff and management that have medical affinity so that beds are interchangeable at these wards (e.g., orthopedic surgery with traumatology or nephrology with endocrinology wards).
- Introduce a new management consisting of a physician and nurse manager.

The hospital management decided that the total capacity should not be increased, so all interventions should be achieved without increase in the number of beds and nurses.

4.2.2 Analysis of Possible Interventions

As boarders are a risk to the quality of care, the hospital wanted to minimize the probability of refusals at the medically preferred ward. As presented in Sect. 3, queueing models dominate strategic and tactical analyses. We therefore chose to

[4]This case study was conducted by the author, among others, A.J.(Thomas) Schneider in the role of healthcare logistics advisor.

model multiple scenarios of assignments of patient groups among wards as an $M/G/s/s$ queue. Based on [13], we analyzed each scenario (e.g., the patient load from selected medical specialties at a ward or merged wards) on two performance measures: (1) what is the blocking probability given an occupancy rate of 85%, and (2) what is the occupancy rate given with a blocking probability of 5%? Based on these insights we redistributed the patient groups. Furthermore, we developed a simulation model to analyze the flow of acute admissions via the acute medical unit (AMU). Patients stay at the AMU for at most 2 days. When further treatment is needed, they are transferred to the inpatient ward of their medical specialty. From this analysis it appeared that solely introducing an AMU would not solve the problem of emergency admission refusals. We therefore analyzed the number of beds at wards that should be allocated to minimize the number of refusals. We also showed the effect of bed shortage at wards (e.g., finally resulting in an overcrowded AMU and emergency department). To prevent flow congestion, we analyzed scenarios with different numbers of beds dedicated for these transfers at each inpatient ward. We used two heuristics to find the best number of beds for each ward or merged wards. This simulation study was executed by a healthcare logistics consultant and ward management (nurse manager and medical manager of the AMU).

4.2.3 Choosing an Intervention

The actual re-dimensioning of wards and redistribution of medical specialties over these wards were completely based on the queueing model results. This required significant effort to convince the stakeholders of the reliability of the model outcomes. We used pseudomized admission data of 2013 and 2014 as input for the model and invested several weeks in discussing the assumptions of the model, results, and data. This is an important step when using queueing models in practice. Although queueing models are based on straightforward formulas, it can be challenging for stakeholders to interpret. For acceptance, the model should be thoroughly discussed and not used as a "black box." We planned special sessions with each medical specialty to discuss the data, showing first the current patient load at each ward. We showed individual patients records to medical staff in terms of admission and discharge dates. Next, we discussed the model assumptions and the used input and key performance indicators. Lastly, we showed the proposed redistribution of patient groups and the effect on the bed usage. Taking time to present the model and answering questions from all stakeholders in both plenary and individual settings, we finally convinced most stakeholders of the proposed redistribution and re-dimensioning.

4.2.4 After Intervention

After the interventions have been completed, the hospital still struggles to create a schedule for patients for at the weekday ward such that this ward can be closed during the weekends. Furthermore, on a later point in time, some medical specialties are again redistributed among wards resulting from new insights to organize wards according to the new strategy of the hospital that implied more thematically care (e.g., oncology care and transplantation care). Again, a queueing model was used for this new distribution of medical specialties.

4.2.5 Lessons Learned

A pitfall in this type of analyses is the requests for more up-to-date data. Given the size of the project, we needed 6 months to discuss the analyses with all stakeholders. During this time, many things can change, and therefore some stakeholders requested a new analysis with up-to-date input data such that the model would be more reliable (e.g., real-time hype). These requests delayed the project by a year. In the end, we did not update the data any further and reasoned with stakeholders that we used highly aggregated data over a long time horizon, which means that mainly trends and/or strategic decisions are detectable in the results. We also showed the added value of merging wards given the performance measures. In practice merging wards in university hospitals has major implications for nursing staff. As nursing staff is highly trained for specific treatments and specialties, they now have to be trained in other and/or more fields of medical expertise as more patient types can be placed at merged wards.

Using a simulation model for the introduction of the AMU, we were able to show stakeholders how the emergency admissions process evolved over time [42]. This visual representation and the implementation of tailored process characteristics significantly contributed in convincing stakeholders. Therefore achieving consensus was easier compared to the use of queueing models for re-dimensioning and redistribution of wards. This was also reflected in the time needed to convince all stakeholders (2 months). The simulation model was used as a tactical tool in the planning and control cycle; in the model we updated the distribution of dedicated beds in each quarter using data with a rolling horizon (e.g., adding the last quarter and deleting the first quarter of the data).

The success of this project was a result of clear sense of urgency throughout the organization to become future-proof. Staff at the operational level dealt with the consequences of the suboptimal distribution of beds and small units on a daily basis. The determination of the board of directors and the persuasiveness of the management convinced all stakeholders to let go of strict bed allocation policies, resulting in larger units. Also the use of operations research models gave the management a safe environment to experiment with new bed distribution and their key performance indicators.

4.3 Case Study III: Bed Assignment Optimization

The Massachusetts General Hospital (MGH, USA) deals with operational bed occupancies between 95% and 100%.[5] As a consequence, patients generally wait long before an inpatient bed is available upon admission or transfer. This results in flow congestion at the postanesthesia care unit (PACU) and the emergency department (ED). Especially for emergency patients long waiting times increase risks. The state of Massachusetts therefore has the policy "Code Help," requiring hospitals to move all admitted inpatients out of the ED within a 30-min period after the ED's maximum occupancy – influenced by the number of patients present and their acuity – is reached or exceeded. Activating Code Help causes the hospital to prioritize moving patients out of the ED, which results in delaying bed assignments for patients from other areas of the hospital, potentially requiring cancellation of elective surgeries and other activities. The consequences of Code Help require significant management attention and can affect hospital operations for several days. In 2015, notifications that the hospital was approaching or had reached Code Help frequently occurred multiple times per week.

4.3.1 Project Organization

The continuous lockdown gave rise to a hospital-wide redesign of admission scheduling. Under supervision of the CEO, a project was initiated that consisted of the head of the perioperative department, the head and a bed manager from the admitting department, the nurse managers and resource nurses of several clinical units, a professor, a postdoctoral fellow, a graduate student in healthcare operations research, and two advisors from the department of process improvement of the MGH.

4.3.2 Analysis of Possible Interventions

Before the intervention, elective surgical same-day admits were preassigned to beds that were occupied with patients who were to be discharged on that day. This was done to guarantee a continual patient flow from the PACU to inpatient units; by reserving a bed for a surgical patient, the bed could not be taken by a patient from the ED or another department. However, the exact timing of discharges was unknown and uncertain. As a consequence, patients were frequently waiting for their preassigned beds, while simultaneously other beds were waiting for their preassigned patients. Data analysis showed that the average time patients waited for a bed ranged between 2.5 and 26.9 h, while a subset of four surgical inpatient

[5]This case study was conducted by, among others, our CHOIR colleague Aleida Braaksma [7] in the role of postdocoral fellow.

units (129 beds) experienced a total bed-wait-for-patient time of 11,181 h or 466 bed-days in 2015.

To alleviate this problem, a simulation model was developed to investigate the effects of multiple interventions, among which is a just-in-time (JIT) bed assignment strategy. This strategy only assigns patients to empty beds just before the moment they are medically ready, and therefore beds could not be preassigned to patients. A second intervention that was investigated was virtually pooling the capacity of two surgical wards, as they were clinically similar. This intervention implies that admissions are not preassigned to one of the wards, but the ward is decided upon the moment the patient is assigned to a bed.

The input for the simulation model is 1 year of hospital data including timestamps for admission and discharge. From the data, empirical distributions were determined for, for example, bed cleaning duration and patient transportation time. The model was made more realistic by implementing bed closures according to the hospital data (e.g., due to staffing shortages) and the hospital's policy with respect to gender and infection precautions in semiprivate rooms. Additionally, the model improved patient cohorting by occasionally swapping a patient from one room to another, mimicking the policy that was used in practice.

4.3.3 After Intervention

Based on the simulation results and 2 earlier projects by graduate students in healthcare operations research, the hospital implemented the JIT and pooling policy on 12 surgical inpatient units. In the 5 months post-implementation, the average patient wait time for bed decreased by 18.1% for ED-to-floor transfers ($P < 0.001$), by 30.5% for PACU-to-floor transfers ($P < 0.001$), and by 10.0% for ICU-to-floor transfers ($P < 0.05$). As a consequence, patients receive their required care earlier, which improves the quality of care. Additionally, the intervention resulted in a smoothed workload for nurses and bed cleaners and less congestion in the ED and PACU. Another positive side effect is an increased focus on patient flow: due to the JIT, nurses wonder why a bed is empty for a long time, which may speed up for example handoffs and transportation.

4.3.4 Lessons Learned

For physicians and nurses, simulation is relatively easy to understand compared to mathematical modeling. Therefore, the project team was convinced the intervention would have a positive effect in practice. The determination of involved clinical leadership (e.g., the head of the perioperative department) was also key for success. In the first days of implementation, nurses were sometimes skeptical about the intervention. The project team leaders showed empathy for the struggles related to the new situation while simultaneously encouraging nurses to stay put. Daily short evaluation meetings gave the opportunity to quickly react to unforeseen side effects

of or negative sentiments around the intervention, before these could evolve into larger problems. The new policies resulted in more stable admission and discharge rates throughout each day. These more stable rates were also noticed by the bed cleaning department as their peaks in workload were reduced. This was therefore also seen as a nice (unforeseen) result of the project.

4.4 Increasing Impact in Practice

In this section we provide our viewpoint on how OR researchers can increase the likelihood of results being implemented in practice. In our previous work van de Vrugt et al. [45] we provided conditions that support a successful implementation in practice.

One of the most important contributions to a successful implementation is the involvement of one or multiple so-called clinical leaders, who are important medical stakeholders in the process and have the respect of all their colleagues. This leader should be able to speak on behalf of his/her colleagues and should discuss the project often with peers. As a researcher, you should earn the trust of these clinical leaders such that they are convinced about the methodological approach and proposed interventions. Ultimately these clinical leaders can (and should) convince other colleagues.

In any model, assumptions are necessary for tractability. Some assumptions may in reality be too unrealistic to be of practical relevance. Therefore, in all our projects we start with one or several observation rounds, in which we study the process in reality, get familiar with the practitioners and their decisions, get to know the assumptions that are important to strictly hold, and see what flexibility and stochasticity are present. Letting practitioners draw a typical patient process is often not accurate enough to obtain all modeling assumptions. Moreover, seeing the outcomes of the process in the data does not mean that the preferred medical process was followed. The time that is invested in making the assumptions and relations in the model more realistic will significantly reduce the time spent on data analysis. Additionally, making the model more realistic will increase the likelihood of adaptation in practice.

Furthermore, to increase to likelihood of implementation, a researcher should be able to convey how the model works to healthcare practitioners and thereby earn the trust of practitioners. Expectation management is very important; practitioners should know what the model can and cannot do. Often when the mathematics behind the modeling approach becomes less complex, the results become easier to grasp and trust by practitioners. A graphical simulation model, in which practitioners should recognize their everyday working routines, is an easy way to earn trust and convince practitioners. After the project is completed for the practitioners, a researcher can decide to continue with thoroughly investigating the problem or extending the model, to make the modeling approach interesting enough to publish

in a OR journal. Otherwise, or perhaps simultaneously, publishing together with the practitioners in medical journals may be considered.

In academia we are used to thoroughly investigate every interesting feature of a model before we present our results. In practice, an iterative process will be more effective; first mimic the current process and let practitioners check it (and repeat this if necessary), and second iteratively investigate a few scenarios and discuss them with practitioners. The most promising for implementation is when the stakeholders actively participate in the iterative process by proposing the interventions to be investigated. Additionally, presenting the results in easy-to-grasp graphics will increase their impact; checking the results with the involved clinical leader before they are presented to all practitioners improves adapting the presentation to the audience. One risk of this iterative process is that the project never ends as more and more scenarios are investigated. This risk can be avoided by setting clear performance targets first in the project and by keeping a strict project schedule.

Discussing possible interventions can be challenging because desired outcomes are often based on extreme incidents. Exploring interventions mathematically and thus rationally often simplifies the discussions significantly. Conveying the chosen intervention to colleagues becomes easier for the clinical leaders, as the decision was based on rational arguments.

5 The Future State of OR for Wards

As a result of continuing innovations, we expect the trend of decreasing LOS to continue. Decreasing LOS results in increasing turnover. With more admissions per bed and with the same turnover time for beds to become available again, the total downtime of beds (time between a discharge and the new admission on one bed) will increase. So minimizing the downtime will be a subject for future research, for example, by looking at the planning of bed cleaning or the number of spare beds at a ward.

Another direction for future research will be the use of big data for trend predictions and incorporating these predictions in the strategic admission planning. This implies that the number of beds at a ward can vary throughout the year. An example is the prediction of the yearly influenza epidemic and allocating more beds for pneumonia diseases during the winter.

Nowadays most hospitals have an electronic patient record, and this (anonymized) data can be used to improve planning and scheduling. Descriptive models (e.g., machine learning) can classify and/or cluster these data. The results of descriptive models can be used as input for operations research models. By combining these techniques, the reliability of results will increase. An example for wards is the prediction of a patient's LOS, based on multiple (preferably routinely collected) patient characteristics.

Finally, we see trends into more regional collaborations between hospitals for specific patient groups. For example, elderly patients may live in a nursing home but may still be treated by hospital doctors. Patients are treated and monitored outside the hospital for as much as possible, by nursing homes, home healthcare, or general practitioners. Using stochastic network techniques, these collaborations can be optimized, in order to, for example, determine adequate capacity levels at all network locations.

Acknowledgments We are sincerely grateful to our colleagues, Aleida Braaksma and Maartje E. Zonderland, for their valuable input.

References

1. Ilgin Acar and Steven E. Butt. Modeling nurse-patient assignments considering patient acuity and travel distance metrics. *Journal of Biomedical Informatics*, 64:192–206, 2016. ISSN 15320464. https://doi.org/10.1016/j.jbi.2016.10.006. URL http://www.sciencedirect.com/science/article/pii/S1532046416301393.
2. Christiane Barz and Kumar Rajaram. Elective Patient Admission and Scheduling under Multiple Resource Constraints. *Production and Operations Management*, 24(12):1907–1930, 2015. ISSN 19375956. https://doi.org/10.1111/poms.12395.
3. R. Bekker and A. M. de Bruin. Time-dependent analysis for refused admissions in clinical wards. *Annals of Operations Research*, 178(1):45–65, 2010. ISSN 02545330. https://doi.org/10.1007/s10479-009-0570-z.
4. René Bekker and Paulien M. Koeleman. Scheduling admissions and reducing variability in bed demand. *Health Care Management Science*, 14(3):237–249, 2011. ISSN 13869620. https://doi.org/10.1007/s10729-011-9163-x.
5. Rym Ben Bachouch, Alain Guinet, and Sonia Hajri-Gabouj. An integer linear model for hospital bed planning. *International Journal of Production Economics*, 140(2):833–843, 2012. ISSN 09255273. https://doi.org/10.1016/j.ijpe.2012.07.023.
6. A Braaksma, J Deglise-Hawkinson, B T Denton, M P Van Oyen, R J Boucherie, and M R K Mes. Online appointment scheduling with different urgencies and appointment lengths. 2014.
7. Aleida Braaksma, Elizabeth Ugarph, Retsef Levi, Ana Cecilia Zenteno, Bethany J Daily, Benjamin Orcutt, and Peter F Dunn. Just-in-time Bed Assignment Improves Surgical Patient Flow. 2018.
8. James R. Broyles, Jeffery K. Cochran, and Douglas C. Montgomery. A statistical Markov chain approximation of transient hospital inpatient inventory. *European Journal of Operational Research*, 207(3):1645–1657, 2010. ISSN 03772217. https://doi.org/10.1016/j.ejor.2010.06.021. URL http://www.sciencedirect.com/science/article/pii/S0377221710004273.
9. Carri W. Chan, Jing Dong, and Linda V. Green. Queues with Time-Varying Arrivals and Inspections with Applications to Hospital Discharge Policies. *Operations Research*, 65(2):469–495, 2016. ISSN 0030-364X. https://doi.org/10.1287/opre.2016.1536.
10. Elizabeth A. Crawford, Pratik J. Parikh, Nan Kong, and Charuhas V. Thakar. Analyzing discharge strategies during acute care: A discrete-event simulation study. *Medical Decision Making*, 34(2):231–241, 2014. ISSN 0272989X. https://doi.org/10.1177/0272989X13503500.
11. J. Dai and Pengyi Shi. Inpatient Overflow: An Approximate Dynamic Programming Approach. *Ssrn*, Available, 2017. https://doi.org/10.2139/ssrn.2924208.
12. Theodore Eugene Day, Albert Chi, Matthew Harris Rutberg, Ashley J. Zahm, Victoria M. Otarola, Jeffrey M. Feldman, and Caroline A. Pasquariello. Addressing the variation of post-surgical inpatient census with computer simulation. *Pediatric Surgery International*, 30(4):449–456, 2014. ISSN 14379813. https://doi.org/10.1007/s00383-014-3475-0.

13. A. M. de Bruin, R. Bekker, L. van Zanten, and G. M. Koole. Dimensioning hospital wards using the Erlang loss model. *Annals of Operations Research*, 178(1):23–43, 2010. ISSN 02545330. https://doi.org/10.1007/s10479-009-0647-8.
14. Andreas Fügener, Erwin W. Hans, Rainer Kolisch, Nikky Kortbeek, and Peter T. Vanberkel. Master surgery scheduling with consideration of multiple downstream units. *European Journal of Operational Research*, 239(1):227–236, 2014. ISSN 03772217. https://doi.org/10.1016/j.ejor.2014.05.009. URL http://www.sciencedirect.com/science/article/pii/S0377221714004160.
15. Steve Gallivan and Martin Utley. A technical note concerning emergency bed demand. *Health Care Management Science*, 14(3):250–252, 2011. ISSN 13869620. https://doi.org/10.1007/s10729-011-9158-7. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-80051895412&partnerID=40&md5=7c0c5179a87cf96ca96a7d29e659a762.
16. Cheng Gao, Abel N Kho, Catherine Ivory, Sarah Osmundson, Bradley A Malin, and You Chen. Predicting Length of Stay for Obstetric Patients via Electronic Medical Records. *Studies in health technology and informatics*, 245:1019–1023, 2017. ISSN 1879-8365. URL http://www.ncbi.nlm.nih.gov/pubmed/29295255%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5860660.
17. Daniel Gartner and Rainer Kolisch. Scheduling the hospital-wide flow of elective patients. *European Journal of Operational Research*, 233(3):689–699, 2014. ISSN 03772217. https://doi.org/10.1016/j.ejor.2013.08.026. URL http://www.sciencedirect.com/science/article/pii/S0377221713006917.
18. S C Graves. Manufacturing planning and control systems. In Resende M Pardalos P., editor, *Handbook of Applied Optimization*, pages 728–746. Oxford University Press, New York, US, 2002. https://doi.org/10.1080/09537289008919307. URL http://web.mit.edu/sgraves/www/ProdPlanCh.PDF.
19. J. D. Griffiths, V. Knight, and I. Komenda. Bed management in a Critical Care Unit. *IMA Journal of Management Mathematics*, 24(2):137–153, 2013. ISSN 14716798. https://doi.org/10.1093/imaman/dpr028. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84877123138&partnerID=40&md5=88190866718e68059cd1e1dbfeeeb7cf.
20. Rosita Guido, Maria Carmela Groccia, and Domenico Conforti. An efficient matheuristic for offline patient-to-bed assignment problems. *European Journal of Operational Research*, 268(2):486–503, 2018. ISSN 03772217. https://doi.org/10.1016/j.ejor.2018.02.007. URL http://www.sciencedirect.com/science/article/pii/S0377221718301218.
21. Erwin W. Hans, Mark van Houdenhoven, and Peter J. H." Hulshof. A framework for healthcare planning and control. In Randolph Hall, editor, *Handbook of Healthcare System Scheduling*, pages 303–320. Springer US, Boston, MA, 2012. ISBN 978-1-4614-1734-7. https://doi.org/10.1007/978-1-4614-1734-7_12.
22. Jonathan Helm and Mark P. Van Oyen. Design and Optimization Methods for Elective Hospital Admissions. *Ssrn*, 62(6):1265–1282, 2014. ISSN 0030-364X. https://doi.org/10.2139/ssrn.2437936.
23. Lene Berge Holm, Hilde Lurås, and Fredrik A. Dahl. 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(2):80–89, 2013. ISSN 13865056. https://doi.org/10.1016/j.ijmedinf.2012.05.006. URL http://www.ncbi.nlm.nih.gov/pubmed/22698645.
24. Peter J H Hulshof, Nikky Kortbeek, Richard J Boucherie, Erwin W Hans, and Piet J M Bakker. Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS. *Health Systems*, 1(2):129–175, 2012. ISSN 2047-6965. https://doi.org/10.1057/hs.2012.18.
25. Peter J H Hulshof, Richard J. Boucherie, Erwin W. Hans, and Johann L. Hurink. Tactical resource allocation and elective patient admission planning in care processes. *Health Care Management Science*, 16(2):152–166, jun 2013. ISSN 13869620. https://doi.org/10.1007/s10729-012-9219-6.

26. Peter J.H. Hulshof, Martijn R.K. Mes, Richard J. Boucherie, and Erwin W. Hans. Patient admission planning using Approximate Dynamic Programming. *Flexible Services and Manufacturing Journal*, 28(1-2):30–61, 2016. ISSN 19366590. https://doi.org/10.1007/s10696-015-9219-1.
27. Mark W. Isken, Timothy J. Ward, and Steven J. Littig. An open source software project for obstetrical procedure scheduling and occupancy analysis. *Health Care Management Science*, 14(1):56–73, 2011. ISSN 13869620. https://doi.org/10.1007/s10729-010-9141-8.
28. Saif Kifah and Salwani Abdullah. An adaptive non-linear great deluge algorithm for the patient-admission problem. *Information Sciences*, 295:573–585, 2015. ISSN 00200255. https://doi.org/10.1016/j.ins.2014.10.004. URL http://www.sciencedirect.com/science/article/pii/S0020025514009839.
29. Alexander Kolker. Interdependency of hospital departments and hospital-wide patient flows. In Randolph Hall, editor, *Patient flow*, volume 206 of *International Series in Operations Research & Management Science*, pages 43–63. Springer US, Boston, MA, 2013. ISBN 978-1-4614-9511-6. https://doi.org/10.1007/978-1-4614-9512-3_2.
30. N. Kortbeek, A. Braaksma, C. A.J. Burger, P. J.M. Bakker, and R. J. Boucherie. Flexible nurse staffing based on hourly bed census predictions. *International Journal of Production Economics*, 161(0):167–180, 2015. ISSN 09255273. https://doi.org/10.1016/j.ijpe.2014.12.007. URL http://www.sciencedirect.com/science/article/pii/S0925527314003909.
31. Paolo Landa, Michele Sonnessa, Elena Tànfani, and Angela Testi. A Discrete Event Simulation Model to Support Bed Management. In *Simulation and Modeling Methodologies, Technologies and Applications (SIMULTECH), 2014 International Conference on*, pages 901–912, 2014. https://doi.org/10.5220/0005161809010912. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=7095143&tag=1.
32. X. Li, P. Beullens, D. Jones, and M. Tamiz. An integrated queuing and multi-objective bed allocation model with application to a hospital in China. *Journal of the Operational Research Society*, 60(3):330–338, 2009. ISSN 01605682. https://doi.org/10.1057/palgrave.jors.2602565. http://www.jstor.org/stable/40206742.
33. Nelly Litvak, Marleen van Rijsbergen, Richard J. Boucherie, and Mark van Houdenhoven. Managing the overflow of intensive care. *European Journal of Operational Research*, 185(3):1–16, 2006. ISSN 03772217. https://doi.org/10.1016/j.ejor.2006.08.021. URL http://eprints.eemcs.utwente.nl/3588/01/1768.pdf http://www.sciencedirect.com/science/article/pii/S0377221706005819.
34. Fermín Mallor, Cristina Azcárate, and Julio Barado. Optimal control of ICU patient discharge: from theory to implementation. *Health Care Management Science*, 18(3):234–250, 2015. ISSN 13869620. https://doi.org/10.1007/s10729-015-9320-8.
35. Avishai Mandelbaum, Petar Momčilović, and Yulia Tseytlin. On Fair Routing from Emergency Departments to Hospital Wards: QED Queues with Heterogeneous Servers. *Management Science*, 58(7):1273–1291, 2012. ISSN 0025-1909. https://doi.org/10.1287/mnsc.1110.1491. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84864074121&partnerID=40&md5=5ffd55982922df3ebb1dbc173d99b05b.
36. Navonil Mustafee, Terry Lyons, Paul Rees, Lee Davies, Mark Ramsey, and Michael D. Williams. Planning of bed capacities in specialized and integrated care units: Incorporating bed blockers in a simulation of surgical throughput. *Proceedings – Winter Simulation Conference*, pages 1–12, 2012. ISSN 08917736. https://doi.org/10.1109/WSC.2012.6465102.
37. J. P. Oddoye, D. F. Jones, M. Tamiz, and P. Schmidt. Combining simulation and goal programming for healthcare planning in a medical assessment unit. *European Journal of Operational Research*, 193(1):250–261, 2009. ISSN 03772217. https://doi.org/10.1016/j.ejor.2007.10.029. URL http://www.sciencedirect.com/science/article/pii/S0377221707010478.
38. Canan Pehlivan, Vincent Augusto, Xiaolan Xie, and Catherine Crenn-Hebert. Multi-period capacity planning for maternity facilities in a perinatal network: A queuing and optimization approach. In *IEEE International Conference on Automation Science and Engineering*, pages 137–142, 2012. ISBN 9781467304283. https://doi.org/10.1109/CoASE.2012.6386385.

39. Gilles Pesant. Balancing nursing workload by constraint programming. In Claude-Guy Quimper, editor, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9676, pages 294–302. Springer International Publishing, 2016. ISBN 9783319339535. https://doi.org/10.1007/978-3-319-33954-2_21.
40. Yazan F. Roumani, Yaman Roumani, Joseph K. Nwankpa, and Mohan Tanniru. Classifying readmissions to a cardiac intensive care unit. *Annals of Operations Research*, 263(1-2):429–451, 2018. ISSN 15729338. https://doi.org/10.1007/s10479-016-2350-x.
41. R. Schmidt, S. Geisler, and C. Spreckelsen. Decision support for hospital bed management using adaptable individual length of stay estimations and shared resources. *BMC medical informatics and decision making*, 13(1):3, 2013. ISSN 14726947. doi:Article. URL http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed11&NEWS=N&AN=23289448.
42. A. J. (Thomas) Schneider, P. Luuk Besselink, Maartje E. Zonderland, Richard J. Boucherie, Wilbert B. Van Den Hout, Job Kievit, Paul Bilars, A. Jaap Fogteloo, and Ton J. Rabelink. Allocating emergency beds improves the emergency admission flow. *Journal of Applied Analytics*, 48(4):384–394, 2018. ISSN 1526551X. https://doi.org/10.1287/inte.2018.0951.
43. Mustafa Y. Sir, Bayram Dundar, Linsey M. Barker Steege, and Kalyan S. Pasupathy. Nurse-patient assignment models considering patient acuity metrics and nurses' perceived workload. *Journal of Biomedical Informatics*, 55:237–248, 2015. ISSN 15320464. https://doi.org/10.1016/j.jbi.2015.04.005. URL http://www.sciencedirect.com/science/article/pii/S1532046415000726.
44. Steven Thompson, Manuel Nunez, Robert Garfinkel, and Matthew D. Dean. OR Practice–Efficient Short-Term Allocation and Reallocation of Patients to Floors of a Hospital During Demand Surges. *Operations Research*, 57(2):261–273, 2009. ISSN 0030-364X. https://doi.org/10.1287/opre.1080.0584.
45. N. M. (Maartje) van de Vrugt, A. J. (Thomas) Schneider, Maartje E. Zonderland, David A. Stanford, and Richard J. Boucherie. Operations research for occupancy modeling at hospital wards and its integration into practice. In Cengiz Kahraman and Y Ilker Topcu, editors, *Operations Research Applications in Health Care Management*, volume 262, pages 101–137. Springer US, Boston, MA, 2018. https://doi.org/10.1007/978-3-319-65455-3_5.
46. N. M. van Dijk and N. Kortbeek. Erlang loss bounds for OT-ICU systems. *Queueing Systems*, 63(1):253–280, 2009. ISSN 02570130. https://doi.org/10.1007/s11134-009-9149-2.
47. J. Theresia van Essen, Mark van Houdenhoven, and Johann L. Hurink. Clustering clinical departments for wards to achieve a prespecified blocking probability. *OR Spectrum*, 37(1):243–271, 2015. ISSN 14366304. https://doi.org/10.1007/s00291-014-0368-5.
48. Catharina J. Van Oostveen, Aleida Braaksma, and Hester Vermeulen. Developing and testing a computerized decision support system for nurse-to-patient assignment: A multimethod study. *CIN – Computers Informatics Nursing*, 32(6):276–285, 2014. ISSN 15389774. https://doi.org/10.1097/CIN.0000000000000056. URL https://journals.lww.com/cinjournal/Fulltext/2014/06000/Developing_and_Testing_a_Computerized_Decision.6.aspx.
49. Peter T VanBerkel and John T Blake. A comprehensive simulation for wait time reduction and capacity planning applied in general surgery. *Health care management science*, 10(4):373–85, 2007. ISSN 1386-9620. https://doi.org/10.1007/s10729-007-9035-6. URL http://www.ncbi.nlm.nih.gov/pubmed/18074970.
50. Peter T Vanberkel, Richard J Boucherie, Erwin W Hans, Johann L Hurink, and Nelly Litvak. A Survey of Health Care Models that Encompass Multiple Departments. *International Journal of Health Management and Information*, 1(1):37–69, 2010. URL http://alexandria.tue.nl/repository/books/653840.pdf%0A http://eprints.eemcs.utwente.nl/15762.
51. Wim Vancroonenburg, Federico Della Croce, Dries Goossens, and Frits C R Spieksma. The Red-Blue transportation problem. *European Journal of Operational Research*, 237(3):814–823, 2014. ISSN 03772217. https://doi.org/10.1016/j.ejor.2014.02.055. URL http://www.sciencedirect.com/science/article/pii/S0377221714001908.

52. Wim Vancroonenburg, Patrick De Causmaecker, and Greet Vanden Berghe. A study of decision support models for online patient-to-room assignment planning. *Annals of Operations Research*, 239 (1): 253–271, 2016. ISSN 15729338. https://doi.org/10.1007/s10479-013-1478-1.
53. J. Williams, S. Dumont, J. Parry-Jones, I. Komenda, J. Griffiths, and V. Knight. Mathematical modelling of patient flows to predict critical care capacity required following the merger of two district general hospitals into one. *Anaesthesia*, 70(1):32–40, 2015. ISSN 13652044. https://doi.org/10.1111/anae.12839. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-84916883111&partnerID=40&md5=afe89bf9c01b8ccaa56837830abd7aaf.
54. Muer Yang, Michael J. Fry, and Corey Scurlock. The ICU will see you now: Efficient-equitable admission control policies for a surgical ICU with batch arrivals. *IIE Transactions (Institute of Industrial Engineers)*, 47(6):586–599, 2015. ISSN 15458830. https://doi.org/10.1080/0740817X.2014.955151.
55. Zhu Zhecheng. An online short-term bed occupancy rate prediction procedure based on discrete event simulation. *Journal of Hospital Administration*, 3(4):p37, 2014. ISSN 1927-6990. https://doi.org/10.5430/jha.v3n4p37. URL http://www.sciedupress.com/journal/index.php/jha/article/view/3553.
56. Maartje E. Zonderland and Richard J. Boucherie. Queuing networks in healthcare systems. In R Hall, editor, *International Series in Operations Research and Management Science*, volume 168, book section 9, pages 201–243. Springer, 2012. ISBN 1461417333. https://doi.org/10.1007/978-1-4614-1734-7_9.
57. Maartje E Zonderland, Richard J Boucherie, Michael W Carter, and David A Stanford. Operations Research for Health Care Modeling the effect of short stay units on patient admissions. *Operations Research for Health Care*, 5:21–27, 2015. ISSN 2211-6923. https://doi.org/10.1016/j.orhc.2015.04.001.
58. B. Cardoen, E. Demeulemeester, and J. Beliën. Operating room planning and scheduling: A literature review. European *journal of operational research*, 201 (3): 921–932, 2010.