Chapter 5
Operations Research for Occupancy Modeling at Hospital Wards and Its Integration into Practice

N.M. van de Vrugt, A.J. Schneider, M.E. Zonderland, D.A. Stanford, and R.J. Boucherie

5.1 Introduction

Medical and technological advancements are contributing to increase healthcare expenditures and increase numbers of hospitalized patients (Chernew and Newhouse 2012), while at the same time the length of stay (LoS) for these patients decreases. However, healthcare expenditures are still rising (OECD). Society calls for improved cost effective healthcare delivery, which puts pressure on available

N.M. van de Vrugt (✉)
Centre of Operations Improvement and Research (CHOIR) and Department of Applied Mathematics, University of Twente, Postbox 217, 7500AE Enschede, The Netherlands

Jeroen Bosch Hospital, Henri Dunantstraat 1, 5223 GZ's Hertogenbosch, The Netherlands

Leiden University Medical Center, Leiden, The Netherlands
e-mail: n.m.vandevrugt@utwente.nl

A.J. Schneider
Centre of Operations Improvement and Research (CHOIR) and Department of Applied Mathematics, University of Twente, Postbox 217, 7500AE Enschede, The Netherlands

Leiden University Medical Center, Quality of Care-Institute, Albinusdreef 2, 2333 ZA Leiden, The Netherlands
e-mail: a.j.schneider@lumc.nl

M.E. Zonderland • R.J. Boucherie
Centre of Operations Improvement and Research (CHOIR) and Department of Applied Mathematics, University of Twente, Postbox 217, 7500AE Enschede, The Netherlands

D.A. Stanford
Department of Statistical and Actuarial Sciences, University of Western Ontario, 1151 Richmond Street North, London, ON, Canada N6A 5B7

C. Kahraman, Y.I. Topcu (eds.), *Operations Research Applications in Health Care Management*, International Series in Operations Research & Management Science 262, https://doi.org/10.1007/978-3-319-65455-3_5

financial resources. At present, hospitals tend to focus on process improvement by decreasing unnatural (e.g. self-created) variability and alignment of different services.

During hospitalization, patients spend most of their time in wards. These wards are also referred to as inpatient care facilities, and provide care to hospitalized patients by offering a room, a bed and board. Wards are strongly interrelated with upstream hospital services such as the operating theater and the emergency department. Due to this interrelation it is essential to attain a high efficiency level at hospital wards in order to achieve efficient patient flow. Hospital ward management often aims for bed occupancy rates above 85% in order to maximize throughput, leaving little slack for flow fluctuations which results in refused, deferred and/or rescheduled patients.

Operations Research (OR) can give managerial insights about trade-offs between performance indicators, such as bed occupancy rates and blocked patients. Although OR methods have the potential to lead to large improvements in all sorts of processes, it appears that the cases in which the models and/or results have been actually implemented are sparse. Using OR models, possible interventions can be evaluated in a safe environment, reducing the risk of implementing an intervention that appears to be counter-productive.

OR models may be invoked for different objectives, for example to provide insights or to optimize a certain performance measure and what effects changes in demand or supply have on these performance indicators. The logistically important performance measures for hospital wards are throughput, blocking probability and occupancy. A possible objective in this area could be to determine the optimal capacity, warranting a prespecified maximal blocking probability and minimal occupancy levels.

In this chapter we focus on occupancy modeling. Related topics, not covered in this chapter, are for example optimizing the assignment of patients to beds, and optimizing patients' access times. The bed assignment problem becomes important when, for example, a ward accommodates patients with infectious diseases, or patients that do not share rooms with the opposite sex. A patient's access time is the number of days between the request for an appointment and the appointment itself, and may be improved by optimizing patient admission schedules and/or the operating theater schedule. Additionally, material logistics and facility design problems are outside the scope of this chapter.

Our aim in this review chapter is to guide both researchers and healthcare professionals through the OR concepts which have been applied to hospital wards. We first present the terminology on the different types of wards and performance indicators covered in this chapter. Next, we give an overview of articles where OR techniques are applied to ward related problems, followed by some detailed examples on how to apply these models. We conclude the chapter by looking at the integration of OR models into practice and possibilities for further research. As background information, we provide a brief introduction of OR models in the appendix.

5.2 Hospital Ward Types and Terminology

In this section we introduce the types of hospital wards and performance measures as used in this chapter. Throughout this chapter, we define a hospital ward as follows: an area or unit within a hospital where inpatients with comparable medical conditions are admitted to a bed to receive care. This typically involves staying overnight until their medical condition changes in such a way that the patient either leaves the hospital, or is transferred to a ward with a different level of care. We therefore place the beds associated with the operating theater (OT), the emergency department (ED) or the outpatient clinics outside the scope of this chapter, as they usually temporarily accommodate patients that undergo a (short) treatment.

The logistical performance of wards is generally assessed by three indicators which are related to each other: throughput, blocking probability and occupancy. The exact definitions of these three performance indicators is given in Sect. 5.2.2, after our definitions of different ward types.

5.2.1 Taxonomy

In this section we distinguish different ward types based on logistical characteristics: the type of in- and outflow, typical length of stay (LoS) and resources, and planning problems the wards face. Based on the literature cited in this chapter, we distinguish the following types of wards:

- Intensive Care Unit (ICU)
- Acute Medical Unit (AMU)
- Obstetric ward (OBS)
- Weekday ward (WDW)
- General ward

We describe each type of ward in terms of (logistical) characteristics below and demonstrate why it is a different type of ward. An overview of the differences is summarized in Table 5.1, in which '0' denotes average occurrence, occupancy or costs, and '+' ('−') denotes increased (decreased) compared to average.

Intensive Care Unit (ICU) For this category in our taxonomy we group several ward types with similar logistical characteristics: traditional ICUs, specialized ICUs, and Critical, High or Medium Care Units. Specialized ICUs are, for example, stroke units, cardiac care units and neonatal ICUs. High Care and Medium Care Units are sometimes combined and often referred as step-down units between ICUs and general wards. In the United Kingdom these combined wards are also referred 'Critical Care Units'. The difference between high and medium care is generally the necessity of breathing support. The ICU of a hospital accommodates the most severely ill patients who require constant close monitoring and support from advanced medical equipment and staff (nurses mostly on a 1:1 basis and intensivists which are readily available) (Mallor and Azcarate 2014). In the remainder of this chapter we refer to the ward types discussed in this section as 'ICUs'.

Table 5.1 Summary of characteristics per ward type

	ICU	AMU	WDW	OBS	General
Long LoS	+	−	−	0	+
Short LoS	+	+	+	+	0
Acute admissions	+	+	−	+	+
Elective admissions	+	−	+	+	+
Bed occupancy	+	−	+	−	0
Staff/bed ratio	+	+	−	0	0
Equipment	+	0	−	0	0

Due to the used equipment and available staff the ICU has the highest costs per bed of all hospital wards. An ICU preferably doesn't defer patients, as this would imply serious mortality risks. However, the costs per bed do not allow for a large buffer in the number of available beds. Therefore, ICUs tend to be fully occupied, and discharge the least ill patient when a bed needs to be freed for a newly arriving patient, or cancel an elective procedure at the OT which requires ICU capacity afterwards. Patient typically either have a short LoS or a very long LoS, and arrive from the OT, ED, wards or surrounding hospitals.

Acute Medical Unit (AMU) AMUs lack a uniform definition. We think the following definition covers the best definition of AMUs 'an AMU is a designated hospital ward specifically staffed and equipped to receive medical inpatients presenting with acute medical illness from EDs and outpatient clinics for expedited multidisciplinary and medical specialist assessment, care and treatment for up to a designated period (typically between 24 and 72 h) prior to discharge or transfer to medical wards' (Scott et al. 2009). Often, AMUs serve as a buffer for both the ED and inpatient wards. Since an AMU treats only urgent patients and should alleviate ED congestion, management is more focused on throughput and LoS, and the target utilization of the AMU beds is typically lower compared to general wards. AMUs are also known under synonyms as 'emergency observation and assessment ward, 'acute assessment unit' and 'acute medical assessment units'. The review papers available (Cooke et al. 2003; Scott et al. 2009) provide a comprehensive overview of definitions and concepts for AMUs. The inflow mainly consists of acute patients from the ED, outpatient clinics, surrounding hospitals or General Practitioners.

Weekday Ward (WDW) WDWs are wards admitting patients with an expected LoS between 2 and 5 days, which are usually only open on weekdays (Conforti et al. 2011). WDW-type of hospitals are also sometimes referred to as 'Monday to Friday clinic' or 'Week Hospital'. Most patients at WDWs are elective, and can be transferred to regular wards without any health risks. Only patients with a highly predictable LoS may be admitted, which is why WDWs mostly treat patients for which strict treatment protocols apply. Scheduling patients at a WDW is complicated by each patient's different LoS and urgency level, which implies a deadline by which the patient should be treated. The requirement that the ward should be closed during weekends also complicates patient scheduling. Most admissions arrive directly from home.

Obstetric Ward (OBS) Obstetric and Gynecology wards provide care for women during their pregnancy, during and after labor, and also take care of their newborns (Cochran and Bharti 2006). Additionally, Gynecology wards accommodate women with problems regarding their reproductive organs. The women at these wards often require (brief) surgical intervention, and typically a short hospitalization. Some hospitals group these types of wards under names like 'Birthing Center', 'Maternity Clinic', or 'Women's and Child's Center'. Most patients arrive from home, outpatient clinics or other hospitals.

General Wards General wards in hospitals are often dedicated to a single medical specialty such as Neurology, Geriatrics, or Hematology. As these wards are generally equipped with similar resources and accommodate both acute and elective admissions which and differ in LoS, we aggregate these ward types. General wards can either be surgical or medical and some wards, such as psychiatric or geriatric wards, are closed, implying that patients cannot leave the wards without approval. Other wards are equipped with a specific type of resource, such as dialysis machines and heart monitors. The nurse to patient ratio is often 1:5–1:6. Patients with a particular medical specialty are typically not all accommodated in the same ward, but may also be admitted at for example a WDW or an ICU. Patient inflow is mainly formed by referrals of outpatient clinics, ICUs, General Practitioners or other hospitals.

5.2.2 Terminology

In healthcare a concept such as 'occupancy', which may seem simple at first sight, has several different definitions. Different researchers and healthcare practitioners use different definitions of occupancy, which may result in false comparisons when the used definitions are not clearly stated. Therefore, we define the frequently used concepts in the following paragraphs. We first define different concepts of capacity (based on Vissers and Beech 2005), then define the throughput and blocking probability, and conclude this section with the different concepts of occupancy.

Each ward has a certain capacity, which is expressed in terms of the number of patients and their care intensity that the ward can accommodate. The capacity of a ward is measured by the number of beds and nurses, and there are different types of capacity. The physical capacity is the number of beds at the ward. Each nurse can take care of a certain number of patients in parallel (determined by the nurse to patient ratio), which determines the structural available capacity. Additionally, temporary capacity changes can occur; for example bed closures in holiday periods, or beds that are used which are officially not staffed in case of bed shortage. The structural capacity and temporary changes together determine the (average) realized available capacity.

Suppose, in a highly stylized example, that a hospital ward has 15 beds in a certain area. There are always three nurses scheduled to work at the ward, and each nurse can take care of at most four patients at the same time. Each summer and Christmas holidays the ward experiences decreasing patient numbers, and decides to only schedule two nurses. The holiday periods together last 8 weeks. Then, for this ward the physical capacity is 15 beds, and the structural capacity is 3 (nurses) $\times$4 (patients per nurse) $=$ 12 beds. Due to the holidays, each year has 8 weeks in which only eight beds are open, so the average realized capacity is:

$$\frac{8(\text{weeks}) \times 8(\text{beds}) + (52-8)(\text{weeks}) \times 12(\text{beds})}{52(\text{weeks})} \approx 11.4 \text{ beds}.$$

As mentioned in the introduction of this section, the logistical performance of a ward is assessed by three performance indicators: throughput, blocking probability and occupancy. These indicators are all related to each other. The throughput of a ward can be measured as the number of admissions or discharges per time unit. The blocking probability of a ward is the percentage of patients that request a bed at the ward at an instance that there are no available beds:

$$P_b = \frac{\text{Number of patients not accommodated at ward}}{\text{Total number of patients requesting a bed at ward}} \times 100\%. \tag{5.1}$$

Blocked patients are either accommodated in a different ward, or deferred to another hospital.

In contrast to throughput and blocking probability, bed occupancy can be quantified by three definitions: based a on bed census at certain time, based on real LoS or based on the number of hospitalization days. Here we aim to give an overview of the most commonly used definitions.

One of the definitions of bed occupancy includes the bed census measured once a day at a specified point in time, for example every morning at 10:00 am. Then, dividing the average of these measurements by the structural available capacity, the occupancy is:

$$O_{bc}(t) = \frac{\text{average bed census at time } t}{\text{structural available capacity}} \times 100\%. \tag{5.2}$$

Note that for the occupancy it also matters how the capacity of a ward is calculated; in most hospitals the structural available capacity is used. A slightly different occupancy measure is obtained by taking the average of multiple bed census measurements throughout each day, for example each hour; we denote this measure by $\bar{O}_{bc}$. The advantage of taking more measurements is that it will better reflect actual bed usage.

Hospitals may also define the occupancy of a ward as the ratio between the total time patients were in beds at the ward and the total time available:

$$O_{LoS}(T) = \frac{\text{sum of all LoSs for all patients in time period } T}{\text{structural available capacity} \times \text{time period } T} \times 100\%. \quad (5.3)$$

This measure is calculated using admittance and discharge time stamps for a certain measurement period, or by multiplying the average LoS with the number of patients accommodated at the ward. This occupancy measure reflects the actual time the beds are used, but does not incorporate unavailability due to cleaning of beds.

Until recently, it was common in Dutch hospitals to determine the bed occupancy using the hospitalization days declared to the insurance companies:

$$O_{hd}(T) = \frac{\text{sum of hospitalization days for all patients in } T}{\text{structural available capacity} \times \text{length } T} \times 100\%. \quad (5.4)$$

Financial hospitalization days were counted in integers, and could be declared if the patient is in a bed before 8:00 pm and discharged after 7:00 am the next day. This implied that the occupancy could be over 100% as beds can be reused if patients are discharged early in the day and new patients are admitted in the afternoon. A drawback of this measure is that it cannot be used as a targeted occupancy for all ward types. Such a situation would arise in wards in which patients generally stay for only a part of a day so that multiple patients can be served by the same bed on the same day (e.g. gynecology). In this system, these wards should therefore achieve occupancy targets over 100%, while wards at which patients stay much longer (e.g. geriatrics) will suffer severe bed shortages if the occupancy is over 90%.

This is an example of an arrival and discharge process at a ward, in order to illustrate the different concepts of occupancy. Consider a ward with three beds that is empty at the start of our observation period. We choose to observe the ward from 8:00 am on day 1, until 5:00 pm on day 4. In this period the following patients arrive:

	Arrival		Discharge			
	Day	Time	Day	Time	LoS	Hosp. days
Patient 1	1	8:00 am	2	6:00 pm	1.42	2
Patient 2	1	10:00 am	4	8:00 am	2.92	4
Patient 3	1	3:00 pm	2	8:00 am	0.71	2
Patient 4	2	3:00 am	Patient is blocked		–	–
Patient 5	2	9:00 am	3	8:00 am	0.96	2
Patient 6	3	9:00 am	After day 4		1.33	2
Patient 7	4	10:00 am	After day 4		0.29	1

(continued)

In this example, patient 4 is blocked as patients 1, 2 and 3 fill up all available beds and the first patient that is discharged (patient 3) is not discharged before patient 4 arrives. Note that the LoS for patients 6 and 7 in the table is not their exact LoS but only the part until the end of the observation period. The bed census for this ward is depicted in Fig. 5.1. The blocking probability for this time period equals 1/7≈15%. The different occupancy measures are calculated as follows.

The bed census at 10:00 am for day 1 to 4 is 2, 3, 2, and 1, respectively, so the average equals 2. Therefore $O_{bc}(10\text{ am}) = 2/3 \approx 66.7\%$. The average hourly bed census is 2.2, so $\bar{O}_{bc} = 2.2/3 \approx 74.8\%$.

The sum of the LoS for all patients at this ward in this observation period, T, equals 7.63 days. The length of the observation period is 3.38 days. Therefore, $O_{LoS}(T) = 7.63/(3 \times 3.38) \approx 75.3\%$.

The sum of the hospitalization days declared for these patients is 13, and the total number of days in this observation period is four. Therefore, $O_{hd}(T) = 13/(3 \times 4) \approx 108.3\%$.

The occupancy measure with hospitalization days is always higher than the other occupancy measures. The ordering of the remaining concepts of occupancy depends on the ward studied.

Hospital management determines which of the aforementioned occupancy measures is used, and sets the target throughput level for each ward separately. A high occupancy usually results in a high blocking probability (Bailey 1952). Therefore it is important for management to balance these three performance indicators. Adequate targets for the performance indicators depend on many factors, for example: the capacity of a ward, the fraction of admissions that is acute, the

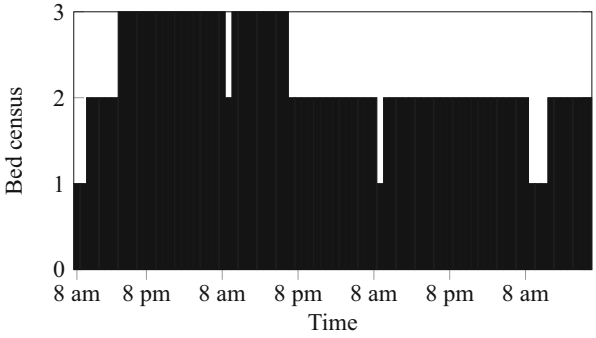

Fig. 5.1 Bed census for example

possibility of deferring admissions, the cost per bed, and the ward layout. Large wards have economies of scale, so a higher bed occupancy can be achieved with a lower blocking probability. If a ward has mostly acute admissions, occupancy targets need to be set lower; elective admissions can be rescheduled in case of bed shortage, while acute admissions cannot. If the deferral of an arriving patient could give rise to life threatening situations (e.g. in case of an intensive care unit), a ward has to lower the target occupancy to produce a lower blocking probability. However, such wards usually have high costs per staffed bed, driving the occupancy targets upwards. Finally, if a ward has many rooms with multiple beds, the bed assignment is less flexible compared to wards with many single bed rooms; if, for example, a patient has an infectious disease he cannot share a room with others. Concluding, it can be said that determining adequate occupancy, blocking probability and throughput targets is a challenging task.

5.3 Ward-Related OR Models

In the previous sections we distinguished different ward types and their logistical similarities and differences. In this section we will review the OR literature for each ward type, emphasizing the main questions or problems the literature tries to solve, the context of problems (e.g. ward type) and the type of models invoked for each paper. An overview and brief explanation of each OR method is given in the Appendix. In Table 5.2 the number of papers found for each ward type and OR model/method is displayed. If a paper invoked multiple OR models, we categorized this paper in all applicable categories. For each ward type, we review the related literature in the following subsections.

Table 5.2 Literature categorized by applied models and ward type

OR model/method	IC	AMU	WDW	OBS	General	**Total**
Algorithms	1	0	0	0	3	**4**
Dynamic Programming	1	0	1	0	0	**2**
Markov processes	4	0	0	2	11	**17**
Mathematical programming	4	2	1	3	6	**16**
Queueing theory	15	2	0	3	16	**36**
Regression	1	0	0	0	1	**2**
Simulation	22	1	0	2	21	**46**
Stochastic models	1	1	0	0	3	**5**
Time series	1	0	0	0	2	**3**
Total	**50**	**6**	**2**	**10**	**63**	**131**

5.3.1 Intensive Care Unit

At the ICU both elective and emergency patients arrive. Emergency patients mostly come from the ED or surrounding hospitals and elective patients mainly arrive after surgery. Since significant costs are involved, management tends to maximize utilization. This results in an increasing number of refusals and/or severe ill patients being transferred from the ICU to high, medium or regular care wards which could lead to situations where quality of care is at stake, and possibly to disruptions in the operating theater schedule. These are also the main problems the literature of this section focuses on: admission and discharge control. See Table 5.3 for an overview of the cited literature in this section.

A queueing model ($M/G/s/s$ queue, see the Appendix for explanation) was used to analyze the total minimum of required ICU beds for burn care for the state of New York (Blair and Lawrence 1981). The authors start by finding the number of beds at an aggregate level given a maximum blocking probability of 5%, and then apply a heuristic allocating these ICU beds among several regional units, while trying to maintain the same blocking probability. This model is extended to analyze an overflow model (Litvak et al. 2008). Here each ICU reserves bed capacity for regional emergency patients, which may be used as overflow beds in a certain region. To approximate the blocking probability of this overflow model the Equivalent Random Method is used, while a simulation model is used to validate the results of this queueing model with historical data. Another modified $M/M/s/s$ model is used to analyze different admission policies and their relation to survival gains (Shmueli et al. 2003). The policies consisted of: (1) the standard first come first served (FCFS) discipline; (2) arrivals are served if and only if a bed is available and the survival gain is greater than an arbitrary threshold value; and (3) arrivals are served if and only if a bed is available and the survival gain threshold value is met, where in this policy the threshold value is depending on the number of beds available. (If fewer beds are available, the threshold value for survival gain will increase.) The results show significant increase in survival gain in both the second and third policy compared to the first policy. The third policy showed only marginal survival gain compared to the second policy, while the number of rejected patients increased significantly. Another application of the $M/M/s/s$ queue is used for analyzing an ICU (McManus et al. 2004). This model is validated with observed data and it is proved that the calculated blocking probabilities from the queueing model were accurate.

Next to queueing, discrete time Markov chains are also applied to ICUs. The authors developed a Markov chain in order to analyze so called bumping (patient transfers from the ICU to free capacity for new arrivals which are more severely ill) (Dobson et al. 2010). Another application is used for the effect of ICU discharge strategies and bed census on patient mortality and total readmission load (patients that are hospitalized shortly after their last admission for the same medical condition) (Chan et al. 2012).

Table 5.3 Literature on ICUs categorized by applied models

OR model/method	References
Algorithms	Blair and Lawrence (1981)
Dynamic programming	Chan et al. (2012)
Markov processes	Blair and Lawrence (1981), Broyles et al. (2010), Dobson et al. (2010), Garg et al. (2010)
Mathematical programming	Kokangul (2008), Mallor and Azcarate (2014), Mallor et al. (2016)
Queueing theory	Blair and Lawrence (1981), de Bruin et al. (2007); De Bruin et al. (2010), Griffiths et al. (2013a), Kim et al. (1999), Litvak et al. (2008), Mallor et al. (2016), McManus et al. (2004), Shmueli et al. (2003), van Dijk and Kortbeek (2009), Williams et al. (2015), Yang et al. (2013), Zonderland and Boucherie (2012)
Regression	Mallor and Azcarate (2014)
Simulation	Bountourelis et al. (2011, 2013), Costa et al. (2003), Davies (1994), Kim et al. (1999), Kim et al. (2000), Kokangul (2008), Kolker (2013), Litvak et al. (2008), Mallor and Azcarate (2014), Mallor et al. (2016), Marcon et al. (2003), Marmor et al. (2013), Masterson et al. (2004), Mustafee et al. (2012), Nguyen et al. (2003), Ridge et al. (1998), Shahani et al. (2008), Sissouras and Moores (1976), Troy and Rosenberg (2009), Yang et al. (2013)
Time series	Garg et al. (2010)

Simulation is also often applied to analyze the required number of ICU beds. In Ridge et al. (1998), Kokangul (2008) and Marmor et al. (2013) they analyze several scenarios, for instance reserving ICU beds for emergency arrivals using simulation. Kim et al. (1999) simulate several ICU arrival processes and compare these results with theoretical results using an $M/M/s$ queue. Based on the simulation model, the authors also determine the blocking probability for the current capacity. Another study (Kim et al. 2000) analyzes several scenarios to minimize the number of elective surgery patients refused at the ICU. The efficient frontier method is used to plot the trade-off between the number of canceled surgeries and the average waiting time per scenario.

Some studies combine several OR techniques to analyze the ICU (Mallor and Azcarate 2014): first, a regression model is proposed for modeling the ICU LoS; second, a comprehensive simulation model is developed for analyzing system behavior and blocking probabilities; and last mathematical programming is used to model the triage problem (which current and arriving patients require ICU capacity the most?) for early or delayed discharges from the ICU depending on high or low utilization of ICU capacity.

When analyzing patient logistics at the ICU, there is a clear distinction between the type of models used and the type of problems solved. Because a significant part of the arrivals at the ICU is unscheduled, queueing theory gives accurate and

representative results. To analyze ICU dynamics, this technique is typically used to achieve general insights about blocking probability, occupancy, ICU capacity, and their mutual trade-offs. Markov chains are used to analyze bed census probabilities and the probability of bumping. Simulation is generally used to analyze multiple scenarios where particular details are involved and/or case-specific dynamics need to be studied.

5.3.2 Acute Medical Unit

The reviews on AMUs mentioned in Sect. 5.2.1 conclude that AMUs may have many advantages, but also that the evidence of economic effectiveness is thin. The AMU 'performance is dependent on good management and availability of diagnostic services', and asserted that there is no proof of cost-effectiveness of AMUs (Cooke et al. 2003). An extensive list of success factors for AMUs is also available (Scott et al. 2009). From an OR perspective, if a hospital does not add beds or staff to its current capacity for opening an AMU, the improved performance reported in the reviews is disputable. The beds assigned to the AMU are taken from other wards, decreasing the benefits of economies of scale and affecting other patients at those wards, and additionally, patients that require inpatient care after their stay at the AMU encounter more process steps than if they would have been admitted directly. Therefore, the effects of opening an AMU cannot be predicted beforehand without the use of appropriate mathematical models. Perhaps partly since AMUs are a relatively new concept, the OR literature with an AMU application is somewhat sparse. In this section, we review this available literature.

Depending on the performance measures of interest and research goals, several models could be applied to AMUs. We describe a goal programming approach used to minimize the delay from ED to AMU, and two different queueing networks to evaluate blocking probability and bed census.

A goal programming approach to determine the required additional resources (beds, doctors and nurses) for each hour of the day to minimize the delay patients experience on an AMU staffed with eight beds, two nurses and three doctors is used (Oddoye et al. 2007). Goal programming is an extension to mathematical programming, in which for each, typically conflicting, objective a target (or goal) is set and deviations from these targets are minimized. In the model, each patient requires a bed, and a specific treatment by a nurse, doctor, or both. A patient is delayed if there are no beds available upon arrival, or if the doctors and nurses are seeing other patients at the moment the patient requires care. For the case studied, the average LoS is 5 h, and the run time of the model equals a day and a half. The conclusion is that only two doctors are required, and a third nurse should be standby in the afternoon and at midnight to cope with peak demand.

In a follow-up study for a larger AMU (currently 58 beds), a simulation study analyzes 14 scenarios with different numbers of beds (Oddoye et al. 2009). Here, each resource type (beds, nurses, and doctors) has its own queue, and patients wait

in these queues until the resource they require is available. Initial targets for each queue length are fed into a goal programming model, together with targets for total LoS and the number of beds. The authors minimize weighted positive deviations from these targets. The model output comprises the resource levels that minimize patients' delay at the AMU, and a trade-off between economic objectives, i.e. higher utilization of resources, and patient- and staff-related objectives is provided.

Another study analyzes a network with one AMU and an aggregated regular ward, in which patients are transferred between the wards if their care requirements change (Utley et al. 2003b). The authors use an infinite server queueing network to determine the probability that the bed occupancy on either ward exceeds a certain number of beds. Based on this probability, they determine the optimal assignment of the available beds to either the AMU or the regular ward. In case the total mean bed occupancy is 85%, and 91% of the patients require acute care, they conclude that 60–65% of the available beds should be designated for acute care.

For a network comprising an ED, two aggregated wards, and an AMU, one study determines the blocking probability by invoking a network of Erlang loss queues in which the AMU both has direct patient arrivals and serves as an overflow ward (Zonderland et al. 2015). They consider both urgent patients (arriving from the ED) and elective patients. The hospital is only allowed to reallocate existing beds from the wards to the AMU. The equivalent random method is used to analyze the network with overflows, since overflow traffic does not follow a Poisson distribution. This method approximates the original network by truncating an infinite server network. The authors conclude that opening an AMU is beneficial for accommodating urgent patients, but the blocking probability for elective patients increases significantly.

The advantage of a simulation or goal programming approach over queueing networks, is that time-dependent arrivals can be incorporated relatively easy. However, the size of the state space in a goal programming model increases with the time horizon considered, and will explode when several departments of realistic sizes are considered. The drawback of simulation models is that they are not easily applied to other hospitals. The advantage of considering infinite server queues is that straightforward formulas for the analysis exist in the literature.

5.3.3 Obstetrics Ward

There are several OR models that have been applied to OBS wards and maternity clinics in the literature. We describe different queueing theory approaches, a simulation model, a discrete time conditional phase type model, and a discrete time Markov model.

In research conducted almost 40 years ago, the bed occupancy at an OBS ward using an infinite server queue is modelled (McClain 1978). The ward may also admit gynecology patients to achieve higher occupancy rates, but those patients are transferred to other wards if an OBS patient has no available bed upon arrival. The

gynecology patients may only be admitted to the OBS ward when the bed census is lower than a certain threshold. They use an infinite server queue to represent the situation where patients are placed in unstaffed beds as a temporary measure when no official beds are available upon arrival. The results are compared for multiple hospitals when including the national guidelines regarding the admittance of gynecology patients to OBS wards, and state which thresholds are best for certain ward sizes.

Another study calculates the probability of delay, e.g. the probability that there is no bed available upon arrival, using an $M/M/s$ queue (Green and Nguyen 2001). Key to this model is that arriving patients who find all beds occupied wait at the clinic until a bed becomes available. During their waiting time, patients are not treated, as their 'service' commences as soon they are placed in a bed. Inputs are the average LoS found in hospital data and different arrival rates. The authors compare the probability of delay for different occupancy targets and different arrival rates.

For a maternity clinic consisting of different wards, including a neonatal ward and ICU, the Queueing Network Analyzer is used (cf. Zonderland and Boucherie 2012, Sect. 2.4.4) to model the bed occupancy (Cochran and Bharti 2006). The authors evaluate all possible bed arrangements among the wards for the peak arrival rate of the clinic. The best arrangements are then evaluated in a system with an inhomogeneous arrival rate in a Discrete Event Simulation. The authors report that the hospital has implemented some of their recommendations, but instead of reassigning beds the hospital chose to add 15 beds to the ward with the highest bed shortage according to the simulation and the queueing model.

To model different types of wards in a network of multiple maternity clinics independent $M/M/s/s$ queues are also used (Pehlivan et al. 2012). The general Erlang loss formulas for the blocking probability are then fed into a Mixed Integer Linear Program to determine strategic bed assignment policies. Each year the clinic may reassign, open and close beds at the wards and clinics, and each decision entails certain costs. The authors incorporate long term planning, since it is undesirable that one year, a ward closes beds and fires nursing staff, while the next year, these beds are reopened and staff are recalled. The objective of the optimization program is to minimize the costs over the decision horizon. One of their conclusions is that efficiency could be gained if resources are transferred among units that experience different demographic changes (increase or decrease in the number of women giving birth).

In an attempt to improve the occupancy rate of an obstetric clinic, one study investigates different scenarios by means of Discrete Event Simulation (Griffin et al. 2012). Inflow and LoS of the model are based on hospital data; patients in the model follow one of the predefined care pathways through the clinic. The authors conclude that the care pathway based approach reflects reality better than a transition probability based approach when they compare the results of both approaches to hospital data. One of the investigated scenarios includes 'swing rooms', which are rooms that can be used by multiple wards of the clinic, but not at the same time. The

clinic implemented the swing rooms, which proved useful for balancing utilization throughout the clinic during bed census peaks.

A discrete time Markov model is developed to mimic a maternity clinic consisting of four wards (Isken et al. 2011). Patients can flow among units, with the routes patients take depending on their type. The authors define eleven patient types and six arrival streams (e.g. natural birth or cesarean), and the LoS has an empirical discrete distribution. All input is derived from hospital data. Since the model assumes infinite capacity, the authors derive the mean and variance of the bed occupancy at the units in case no patients would be deferred to other clinics. These can be used to approximate the bed census by fitting a normal distribution with the same mean and variance. The normal approximation is included in an Integer Linear Programming (ILP) optimization model to optimize the scheduled arrivals at the clinic. Several of the assumptions are validated by means of a Discrete Event Simulation. One of the conclusions is that scheduling some patients on Saturdays smooths the bed census significantly. The authors report that their model has supported multiple clinics in the United States.

The next study focuses more on predicting the LoS of women arriving at a maternity clinic (Harper et al. 2012). The authors define a phase-type distributed LoS for two labor types: spontaneous and scheduled. For both types a decision tree based on patient characteristics, e.g. age and weight, further specifies the LoS parameters. The prediction of the LoS is then included in a simple continuous time Markov model to calculate bed occupancy for the labor ward of the clinic, using a homogeneous arrival rate. The model uses the LoS distribution and transition probabilities that women experience in each phase of labor. The steady state of the model reflects the bed census at different phases, which require different wards at the clinic.

In the literature on OBS wards we found two attempts at increasing bed occupancy, by either admitting non-OBS patients or by using 'swing rooms'. Interestingly, Harper et al. (2012) conclude that the hospital data they obtained does not show a specific time dependent arrival distribution, while others (Cochran and Bharti 2006; Griffin et al. 2012; Isken et al. 2011) do model time dependent arrival rates. Arguably, scheduled arrivals (scheduled cesarean births) likely occur only during office hours, which implies a time dependent arrival rate. Queueing models are more difficult to use in a time dependent system, since the simple formulas for waiting and blocking probability do not hold in a time dependent system. The drawback of using simulation models is that most models are case-specific, applicable only to the clinic they were designed for. However, the advantage of a graphical simulation is that practitioners can easily see the implications of different interventions, which often implies that results of the research are more easily implemented into practice. An advantage of the discrete time Markov models is that these models have the potential to mimic reality better than queueing models, and are still more general than simulation models. However, a drawback could be a rapidly increasing state space for average sized clinics consisting of multiple wards. Others propose an approximation of the bed census by a Normal distribution, and from their simulation results this seems a reasonable assumption (Isken et al. 2011).

5.3.4 Weekday Ward

Although most Dutch hospitals have a WDW and the optimization potential is significant, we were able to find only two references. This may be explained by the lack of capacity issues in these type of wards. Since all patients are elective, they can be scheduled at a time that beds are available, and patients that cannot be admitted will be accommodated on the general ward. Still, we feel that WDWs have a large logistical potential; large efficiency gains can be achieved if the number of beds is adequate and patient scheduling is optimized.

Due to the lack of modeling work on WDWs and the sparsity of scheduling work for this type of ward, we describe below two models for optimizing the patient scheduling that are relevant to the present discussion.

For a 'Monday to Friday' rheumatology clinic, admissions from a waiting list are optimized (Conforti et al. 2011). An introductory meeting determines a patient's medical priority, resource requirement and LoS. LoS is maximally 5 days. Others develop an ILP, in which they decide for each resource the patient requires (e.g. beds, diagnostic tests) at which time slot it should be scheduled, if any (Conforti et al. 2011). Each patient is assigned a weight according to his medical priority and time spend on the waiting list, while the objective is to maximize the weighted number of admissions. The authors conclude that the number of available beds is the bottleneck, and the optimized schedule can accommodate twice the number of patients compared to the schedule which was composed manually.

The last study on WDWs we found considers an online appointment scheduling version of the WDW patient scheduling problem: a patient's request arrives and should be assigned to a date and time immediately, without knowing future patient arrivals (Braaksma et al. 2015). The authors develop an Approximate Dynamic Programming model to obtain the optimal scheduling policy. This technique is often invoked when Dynamic programming models suffer from 'the curse of dimensionality', and includes aggregating the state space and approximating the value function.

5.3.5 General Ward

This section discusses models which are not applied to a specific type of ward. In most of the literature included in this section, general concepts are analyzed that are applicable to many types of wards, or the studies take multiple departments into account. Due to this generalization, most literature discussed in this section focuses on strategic or tactical planning by evaluating capacity dimensioning decisions or predicting demand.

The models for analyzing general concepts of bed census cover a wide range of OR techniques and are applied on different levels. The techniques used in the literature included in this subsection are given in Table 5.4. We will highlight

Table 5.4 Literature on general wards categorized by applied models

OR model/method	References
Algorithms	Best et al. (2015), Holm et al. (2013), van Essen et al. (2015)
Markov processes	Akkerman and Knip (2004), Gorunescu et al. (2002c), Keepers and Harrison (2009), Kusters and Groot (1996), Ramakrishnan et al. (2005), Shonick and Jackson (1973), Swain et al. (1977), Taylor et al. (2000), Utley et al. (2003a, 2005), Vasilakis et al. (2008)
Mathematical programming	Akcali et al. (2006), Bekker and Koeleman (2011), Best et al. (2015), Li et al. (2009), van Essen et al. (2015)
Queueing theory	Bekker and de Bruin (2010), Bekker and Koeleman (2011), Best et al. (2015), De Bruin et al. (2010), Gallivan and Utley (2011), Garrison and Pecina (2015), Gorunescu et al. (2002a,b,c), Green and Nguyen (2001), Griffiths et al. (2013b), Harrison et al. (2005), Li et al. (2009), Vasilakis and El-Darzi (2001), Zonderland and Boucherie (2012)
Regression	Kumar and Mo (2010)
Simulation	Akkerman and Knip (2004), Bagust et al. (1999), Dumas (1985), El-Darzi et al. (1998), Ferreira et al. (2008), Gorunescu et al. (2002c), Gunal and Pidd (2010), Harris (1986), Harrison et al. (2005), Holm et al. (2013), Keepers and Harrison (2009), Kolker (2013), Kumar (2011), Kumar and Mo (2010), Landa et al. (2014), Lapierre et al. (1999), Vanberkel and Blake (2007), Vasilakis and El-Darzi (2001), Vasilakis et al. (2008), Zhu (2011, 2014)
Stochastic models	Kortbeek et al. (2015), Mackay (2001), Vanberkel et al. (2011), Vasilakis et al. (2008)
Time series	Lapierre et al. (1999), Mackay and Lee (2005)

these models and their conclusions below by discussing a selection of the papers in Table 5.4.

A queueing model is used to determine the bed demand at community level, focusing on high occupancy rates, while keeping refusal rates of emergency patients low and waiting lists short (Shonick and Jackson 1973). The bed census is modeled using an infinite server queue incorporating two classes (elective and emergency) of arrival streams. This model elaborates on earlier research applying the infinite server queue by adding a threshold parameter (B) that blocks elective admissions if the occupancy rate is higher than or equal to B, in order to balance the elective and emergency arrival streams. This model provides policy makers useful insights in the relation between bed census, length of the waiting list and emergency refusals. Another queueing model incorporates predictable fluctuations in the average number of arrivals (Bekker and de Bruin 2010). This time-dependent queue, an $M(t)/H/s/s$ model (where $M(t)$ indicates a time-dependent Poisson process, see the Appendix for information on the notation), is evaluated by using approximations based on the infinite server queue. It is shown that daily fluctuations have limited impact on the bed census, whereas weekly patterns do have a significant

impact on both the bed census and the number of refused admissions. Finally, the authors present a method to determine the required number of beds across the week. An $M/PH/s$ queue is used to determine the optimal bed census for a hospital, in which the LoS is phase-type distributed (which is denoted by the abbreviation PH) (Gorunescu et al. 2002a). De Bruin et al. (2010) employ the Erlang loss model ($M/G/s/s$ queue) to relate the blocking probability to the occupancy. Additionally, a broad introduction of various applications of queueing networks in healthcare is also available (Zonderland and Boucherie 2012).

Several papers use a discrete Markovian approach to predict the short term bed census. These predictions are mainly based on the current bed census at day t, the expected elective and emergency admissions, and the expected discharges at day $t+j$. In these models the LoS is often empirically distributed. The census distribution is approximated from their Markov model by a Normal distribution (Utley et al. 2003a), and shows that this relatively easy approximation performs satisfactory when applied to hospital wards. Markov models are also applied to obtain the distribution of the number of patients in each phase of a care pathway, for geriatric (Taylor et al. 2000; Gorunescu et al. 2002c) or stroke patients (Vasilakis et al. 2008), in order to determine the required resources in each phase of the pathway.

Simulation is used by Dumas (1985) to analyze bed allocation and usage policies for all beds in a hospital based on hospitalizations days (e.g. 24 h bed occupancy) per specialty, average daily bed census at a certain time, bed occupancy over a time period, patient misplacements and annual misplaced patient-days. Another simulation analyzes the so called 'winter bed crisis', a yearly bed shortage during mid winter (Vasilakis and El-Darzi 2001). The results show that discharge delays during mid winter were the main reason for high bed census. The following study analyzes waiting times for surgical procedures by means of simulation (Vanberkel and Blake 2007). To balance emergency and elective admissions for the available bed capacity another simulation study was performed (Landa et al. 2014). The last simulation study focusses on the overflows between wards (in which patients are transferred to another ward because the designated ward is fully occupied), and find that the occupancy of wards is a good predictor for the frequency of overflows (Keepers and Harrison 2009).

Time series models are also used to predict bed census demand. An hourly bed sensus prediction was modelled with a time series model (Lapierre et al. 1999). The results are used to reallocate beds between different ward types such as medical, surgical or obstetric. A different but related approach involves the use of mixed exponential equations to obtain the probability distribution of patients being in different phases of their care pathways. In Mackay (2001) and Vasilakis et al. (2008) the model is applied to mimic bed census, allocating emergency admissions on both a regional and hospital level. Results show that this type of model mimics the bed occupancy accurately. The first study analyzes the accuracy of these mixed exponential equations based on a case study, and compare different equations by evaluating the effect of adding more parameters (Mackay and Lee 2005). And the latter study relates the blueprint schedule of the OT, in which each subspecialty

gets a fraction of the available OT time, to the hourly bed census distribution at the postoperative wards (Kortbeek et al. 2015).

A nonlinear mixed integer mathematical programming model is used to (re)allocate the number of available beds among different hospital services over a finite planning horizon (Akcali et al. 2006). The decisions are based on patients' waiting time before admission and budget limits. A similar technique is employed, where integer programming assists in clustering the clinical departments and assigning these clustered departments to available wards (van Essen et al. 2015). These assignments are such that capacity is sufficient to guarantee a maximum blocking probability.

Concluding, the choice for a certain modeling technique depends on the desired output. Queueing theory is suitable for determining the capacity or census distribution of a single ward, preferably with mostly unscheduled patient admissions, when a maximum blocking probability or target occupancy must be achieved. Markov models and time series models are accurate for determining the census distribution or certain percentiles, but might be tedious to analyze as the state space may become large. Simulation models can be developed as detailed or macro-leveled as desired, but are generally suitable for obtaining average performance measures. Mathematical programming can be used to optimize the reallocation of beds to wards.

5.4 Illustrations of OR Model Use

In the previous section we reviewed several OR models applied to different types of wards. In this section we provide several detailed examples of OR models applied to an ICU, OBS, AMU and WDW. All examples are based on hospital data, and illustrate the effectiveness of OR models for certain ward types. The anonymized data used for the examples is obtained from our affiliated hospitals.

5.4.1 ICU Case Study

In this case study we model the bed census of an ICU of a medium-sized Dutch teaching hospital (700 beds in total). The performance measures of interest are the bed occupancy and the probability that the bed census exceeds 40 beds (the current ICU capacity). Queueing models are therefore appropriate to apply to this case study. Hospital data shows that the number of arriving patients per day is Poisson distributed, which was expected as most patients at an ICU are urgent.

Since patients at an ICU require intensive care, deferring patients or letting them wait for a bed is not a viable solution. We therefore model the ICU with the $M/G/\infty$ queue, an infinity capacity queue, so we model the system if all patients would be accepted at the ICU. For tractability, we assume that admissions arrive according to a Poisson process.

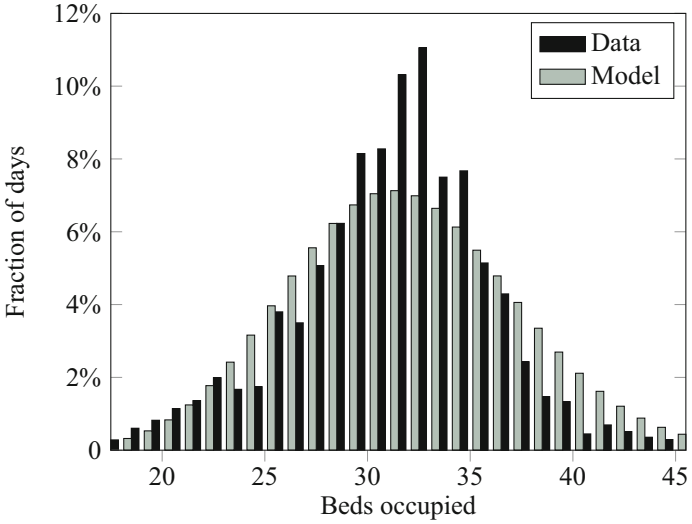

Fig. 5.2 Bed census distribution for the ICU case study

Let λ denote the arrival rate per day, $1/\mu$ the average LoS in days, and $\rho = \lambda\mu$ the load of the system. In an infinite server model the number of patients at the ward at any point in time has a Poisson distribution with parameter ρ (c.f. Winston 2003). Therefore, the probability that n beds are occupied is given by:

$$p_n = \frac{\rho^n}{n!} e^{-\rho}. \tag{5.5}$$

These probabilities are independent of the shape of the LoS distribution of patients, which is convenient for modeling an ICU as the LoS distribution at ICUs typically has a long tail, so the LoS has a high variance. In this case study the variance of the LoS is larger than the average LoS squared.

For the ICU of this case study we find from the data that the average daily arrival rate is 2.18 patients per day, and the average LoS is 14.41 days. Therefore, $\rho = 31.4$. Comparing the bed census from hospital data and the queueing model, see Fig. 5.2, we conclude that the model underestimates the probability of an 'average census' (around 32 beds) and overestimates the probability of most other census values. From the hospital data the probability that the bed census exceeds 40 equals 4.1%, while from the model this is 5.6%.

Note that in the hospital data, the bed census should not exceed the actual capacity, as this represents the realized occupancy. The fact that the census does exceed the capacity at some points in time, might be due to registration errors, for example when a nurse fills out all discharges at the end of the shift instead of the actual time of discharge or only the first ward where a patient is admitted throughout his entire stay is registered in the data (which was the case in this data). Additionally, the actual demand for beds is hard to obtain, since intensivists typically transfer a

relatively healthy patient to another ward in case of bed shortages. This complicates the validity of the model for situations close to full capacity.

With the results of this model, hospital management can gain insight about the current performance of the ICU regarding the probability that bed census exceeds capacity and occupancy. Also, the effect of management decisions can be analyzed with this model, for instance the effects of bed expansion and downscaling, or an increasing average LoS through a different patient mix.

5.4.2 OBS Case Study

In this case study we model the bed census of an OBS with 24 beds. The performance measures of interest are the bed occupancy and the probability that arriving patients have to be deferred because all beds are occupied. A queueing model is therefore an appropriate choice. Hospital data shows that the number of arriving patients per day is Poisson distributed, which was expected as most patients at an OBS are unscheduled. The hospital data shows that the arrival rate is homogeneous over the hours, except for 8:00 am; at that time most of the elective patients at the ward are admitted. For ease of modeling, we assume the arrival rate to be constant throughout the day and week. Again, we assume that admissions arrive according to a Poisson process. The performance measures of interest are best obtained by using a queue with finite capacity: an $M/G/s/s$ queue, also known as the '(Erlang) loss queue'.

Let λ denote the arrival rate per day, $1/\mu$ the average LoS in days, and $\rho = \lambda/\mu$ the average load. In the loss queue the probability that there are n patients present at a ward with capacity s beds, is given by:

$$p_n = \frac{\rho^n/n!}{\sum_{i=0}^{s} \rho^i/i!}. \tag{5.6}$$

These probabilities are independent of the LoS distribution of patients, which is in this case convenient as the LoS distribution at this OBS has a long tail.

For the OBS of this case study we find from the data that the average daily arrival rate is 9.64 patients, and the average LoS is 1.14 days. Therefore, $\rho = 10.96$. Comparing the bed census from hospital data and the queueing model, see Fig. 5.3, we conclude that the model predicts the occupancy quite accurately. The expected number of occupied beds is 10.9 according to the model, and 11.0 according to the hospital data. From the model we can determine that the probability the ward is fully occupied equals 0.025%. As the hospital does not register the number of deferred patients, we cannot verify this result.

The probability of a full ward is useful management information, since then hospital management can determine if the available capacity is still sufficient. Also this model can be used to analyze the effects on blocking probability and occupancy

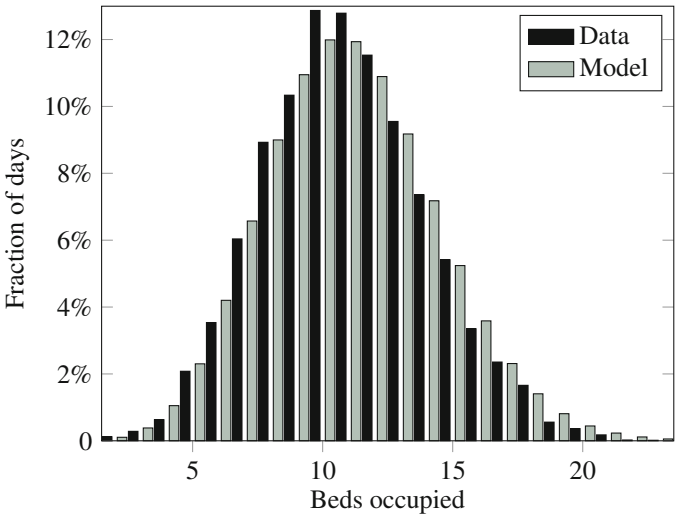

Fig. 5.3 Bed census distribution for the OBS case study

by changing the capacity of the ward. This is in fact easy to do, as there is a simple recursion between the Erlang loss probability for wards differing by one bed (i.e, server).

5.4.3 AMU Case Study

In this case study we consider a medium-sized Dutch teaching hospital (700 beds) that experiences difficulties with allocating urgent medical patients to inpatient beds, especially outside office hours. Typically, medical patients experience a longer ED LoS than surgical patients, partly due to a more complex diagnostic trajectory which involves observation of the patient and waiting until test results are available. As a result, EDs may become congested with this kind of medical patients that are under observation. Therefore, hospital management is considering opening an AMU to support the ED and medical departments. The purpose of the AMU would be faster admittance of ED patients that require observation or short hospitalization.

In preparation of the analysis, the doctors of the hospital have provided a list of diagnoses that can be admitted to the AMU. With this list the number of patients that would be admitted to the AMU if it were opened, can be estimated. Upon AMU discharge, patients either leave the hospital, or are admitted to an appropriate inpatient ward. The doctors agreed that discharges from the AMU would only occur during extended office hours (8:00 am–8:00 pm).

The performance measures of interest are the bed occupancy and blocking probability. Queueing models are therefore appropriate to apply to this case study.

As all patients are urgent and arrival rates at an ED are strongly time-dependent, we model the AMU by means of an Erlang queueing model with time-dependent arrival and service rates: an $M(t)/M(t)/s/s$ queue. Here, M_t denotes a time-dependent Poisson distribution, and s denotes the number of beds at the AMU. In a non-stationary loss queue, the limiting distribution for the number of patients in the system is time-dependent and can only be approximated. Several approximation methods exists, for example the Modified Offered Load (MOL) algorithm (Massey and Whitt 1994). Also in this case for tractability, we assume a Poisson arrival process.

The MOL algorithm approximates the load of the $M(t)/M(t)/s/s$ queue by truncating the state space of an equivalent system with infinite number of servers. Therefore the probability of having n beds occupied at time t at a ward with s beds in total, is given by:

$$P_n(t) \approx \frac{\rho(t)^n/n!}{\sum_{i=0}^{s} \rho(t)^i/i!}, \tag{5.7}$$

with $P_n(t)$ the limiting probability of n patients in the system at time t, and $\rho(t)$ the time-dependent equivalent of $\rho = \lambda/\mu$ satisfying

$$\frac{d}{dt}\rho(t) = \lambda(t) - \mu(t)\rho(t).$$

Here $\lambda(t)$ is the time-dependent arrival rate, and $\mu(t)$ is the time dependent departure rate of the AMU. The MOL approximation provides good results when the system load is moderate. In systems with high load the blocking probability is underestimated.

We obtain the time-dependent limiting probabilities of the number of occupied beds for the hospital by employing the MOL algorithm, and use these probabilities to obtain the expected bed occupancy and blocking probability. We investigate two scenarios: admitting new patients 24 h per day, or only during night time. Input for the model are the time-dependent arrival rate obtained from hospital data, depicted in Fig. 5.4, and time-dependent service rate found in hospital data. The arrival rates are adjusted to reflect the investigated scenarios. Doctors defined the patient types eligible for admitting to the AMU, and data showed that this concerned 26% of the urgent medical patients. For the hospital that commissioned this case study, opening an AMU is not warranted, as the bed occupancy would be low while the blocking probability would be high, as seen in Table 5.5. The number of patients that can be admitted to the AMU is not enough to achieve an acceptable bed occupancy and blocking probability simultaneously. Based on this results, the managers and doctors of this hospital decided not to open an AMU, and investigated other ways to reduce the ED crowding.

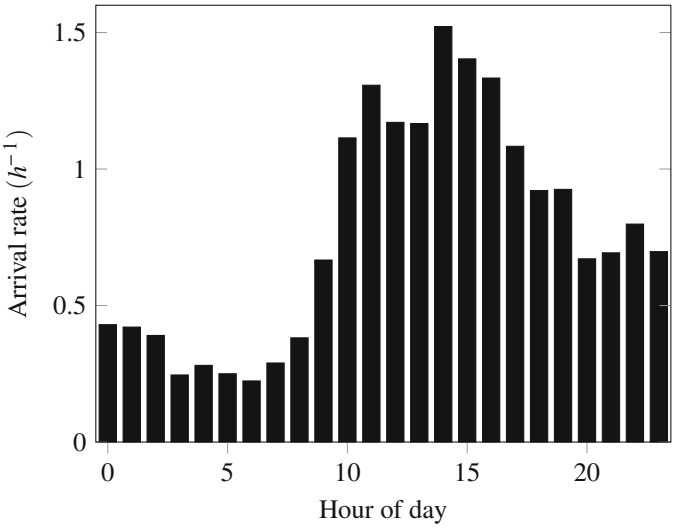

Fig. 5.4 Arrival rate per hour for urgent medical patients

Table 5.5 Results MOL algorithm, 'Occupancy' is average hourly occupancy

24 h/day admissions			Only admissions during nights		
#beds	Occupancy	Block. prob.	#beds	Occupancy	Block. prob.
5	82%	46%	3	61%	37%
8	73%	20%	4	55%	21%
10	66%	9%	6	43%	5%
12	58%	3%			

5.4.4 WDW Case Study

In this case study we optimize patient admissions to a WDW, aiming to minimize the number of beds necessary to accommodate all patients at the ward. The desired model output is both the minimum number of beds required, and a cyclic blueprint patient admission schedule. This blueprint schedule specifies for each day of the week how many of each patient type may be admitted to the ward, and there should be at least enough capacity to accommodate the average number of arriving patients.

The WDW of this case study primarily accommodates outpatients (91%), who do not stay overnight, but typically sleep off their anesthetics after a simple surgery. We assume that the cycle length is 110 h, from Monday 7:30 am until Friday 9:30 pm, as the ward closes in the weekend. Note that in this hospital, patients may be discharged after office hours. We aggregate the possible diagnoses at the WDW according to their LoS, and obtain the patient types from hospital data as given in Table 5.6.

Table 5.6 Patient types at WDW case study

Type description	LoS (h)	Av.no. patients/week
LoS < 0.5 day	5	72
0.5 day < LoS < 1 day	24	47
1 day < LoS < 1.5 day	36	8
1.5 day < LoS < 2 days	48	3

Table 5.7 Possible admission patterns WDW case study, with per LoS type the number of patients in each pattern

Pattern → LoS ↓	1	2	3	4	5	6	7	8	9	10
5	2	2	2	15		5	10	6	8	5
24		1	4		1				1	3
36		2			1	1		2	1	
48	2				1	1	1			
Sum LoS	106	106	106	75	108	109	98	102	100	97

We use an integer linear mathematical programming (ILP) model to solve this problem, which is known as the 'Cutting stock problem' (Roelofs and Bisschop 2012). We assume each bed at the WDW is available for 110 h. We define possible 'admission patterns': a combination of patients that could be placed consecutively in one bed within the opening hours of the WDW. For this case study we manually defined these patterns (see Table 5.7) as implementing too many patterns would not be practical. Note that for patterns 4, 7, 9, and 10, more patients with LoSs of 5 h could be added; we assumed the maximum number of admissions per bed per day is three, to avoid admitting patients outside office hours. The mathematical program determines the minimum number of beds necessary to accommodate all patients.

For the ILP model we need to define sets, parameters, variables, constraints, and an objective. Let $p = 1, \ldots, 11$ be the set of patterns, and $t = 1, 2, 3, 4$ the set of patient types. The parameters of the model are the demand for each patient type, D_t, and the number of patients of each type in each pattern, A_{tp}. D_t is defined in the last column of Table 5.6, and A_{tp} in Table 5.7. Define the decision variables of the ILP by x_p, the number of beds with admission pattern p. The objective is to minimize the sum of all x_p, while the constraints should reflect that all patients can be accommodated. The ILP is given by:

$$\min \sum_p x_p \tag{5.8}$$

subject to :

$$\sum_p A_{tp} x_p \geq D_t \qquad \text{for all } t \tag{5.9}$$

$$x_p \text{ integer for all } p. \tag{5.10}$$

Table 5.8 Solution of the ILP for the WDW case study

Pattern	2	3	5	9	10	11
No. beds	2	6	3	1	3	4

The ILP can be solved with many commercially available solvers, and we chose to use Microsoft Excel®. The patterns that should be used at least once and the total number of required beds are given in Table 5.8. Using this schedule the department has slots for 87 patients with LoS maximally 5 h, and precisely enough slots for the average number of the other patient types. In total 19 beds should be enough to accommodate all patients at the WDW.

The patterns are quite flexible to use in practice as the exact sequence of the patients is not specified. During admission scheduling, the WDW has to take into account that the patients should be discharged before the ward closes on Fridays.

5.5 Implemented OR Results

There exist many papers on OR models relating to different types of wards. It appears that the bed census and/or occupancy can be modeled quite accurately. However, actual use of the models in practice seems scarce; only a few of the articles reviewed for this chapter report on actual implementation results, or use of the models in practice. A widely used quote is: 'the final test of a theory is its capacity to solve the problems which originated it' (Dantzig 1963). In this section we report on the problems faced with while implementing research results, and the lessons learned from the implemented research included in this chapter.

The most important lesson from the literature is that all stakeholders (not necessarily only the problem owners) should be involved throughout the entire process to increase the likelihood of implementation (Harper and Pitt 2004; Cochran and Bharti 2006; Dumas 1985; Harper and Shahani 2002; Harris 1986; Troy and Rosenberg 2009). In the phase of defining the problem, the stakeholders determine the scope of the research, relevant performance measures, and the type of output desired, for example a new admission schedule or a decision support system. When data needs to be collected for the project, stakeholders are important for retrieving data, defining the in- and exclusion criteria, and the validation of the data. Throughout the modeling phase of the project, the stakeholders are involved in several iterations of presenting and discussing preliminary results. In the last research phase, stakeholders and/or hospital management have to decide on the recommendations they want to implement, before the actual implementation can begin.

Model input determines to a large extent the outcome and the acceptance of the results. On several occasions the already available hospital data appeared to be insufficient to provide all necessary input for the models, or the database was incomplete (Kusters and Groot 1996; Lapierre et al. 1999). Hospital data is often

inconsistent or partly missing across different databases; financial data does not always match (raw) admission and discharge data. Depending on the goals of the research, different databases may be used. Even in times of increasing use of technology, we cannot trust the data to reflect reality completely. The entry of admission and discharge data, for example, is in many hospitals still a manual task, often performed when nurses have relatively low workload or at the end of a shift. Additionally, it is important to realize that all hospital data is the *realized* process and most hospitals do not register deferred or denied patients, so actual patient demand is often hard to obtain. Knowing the ins and outs of the healthcare process is also essential in reading the data; for example, for an ICU, the LoS is affected by the bed occupancy since intensivists often transfer the healthiest patient to free a bed for a new patient in case all beds are occupied. A careful sensitivity analysis should be performed to ensure that the best possible scenario for implementation is included in the analysis.

Even when the analysis shows that one of the investigated scenarios is clearly superior to the others, a hospital may decide to implement (slightly) different interventions than those recommended. Discussions during projects often stimulate hospital staff and management to search for further possibilities for improvements (Dumas 1985; Griffin et al. 2012). During the project, a thorough robustness analysis should be performed to ensure that modified recommendations also improve the hospital's processes, and to avoid undesired outcomes and side-effects of the interventions. If possible, the interventions the hospital chooses to implement should be evaluated using the developed models.

There are two types in the implementation of research results: some hospitals implement the model, and some implement the management decision based on the results of the model. When a hospital is using the actual model, a researcher or third party develops a decision support tool that can either be included in the hospital's current software or used separately. The tool should match its user specific settings, or be flexible enough to be adapted to them (Kusters and Groot 1996; Swain et al. 1977). Additionally, users should be trained and supported in working with the model to ensure the continuation of the model's use (Harper and Shahani 2002; Swain et al. 1977).

In some projects measuring the effects of the implementations may be difficult. Hospitals may decide to implement many different interventions at the same time (Griffin et al. 2012; Kusters and Groot 1996), making it impossible to distinguish the precise impact of a particular intervention. The environment in which a ward operates may change, for example when two hospitals merge or the hospital districts are redefined (Holm et al. 2013). For prospective studies it may be unethical to measure the effect of the intervention via a randomized controlled trial, for instance: if opening an AMU appears to be the best scenario for patients, a hospital cannot set up an experiment in which one group is treated in an AMU and a control group is not. Additionally, it may also be too costly or complicated to operate a process in two different ways in parallel. Another difficulty analyzing the practical effects of an implementation is the implication of default behavior by stakeholders in models. Analyzing a system or a population, models optimize the

overall performance, while, for instance, care professionals do not act on a system level but act on individual patients. So the best option for an individual patient could be suboptimal (or even worse) for the system. Therefore the results of a modeling exercise should always be accompanied by its implications for practice. When measuring the interventions' practical effects, one should take the behavior of individuals into account.

In summary, the stakeholders play a significant role in the likelihood of implementation. Additionally, researchers should be thorough in their data collection, sensitivity and robustness analyses, and implementation support. Additional information on project life cycles for general healthcare applications is found in Harper and Pitt (2004).

5.6 Challenges and Directions for Further Research

In this chapter we have discussed various OR techniques applied to different types of wards. We elaborated on to what extent these models are implemented into practice. Some models can be applied to more than one ward type and are often used in literature. We will summarize these general models, and discuss implementation and opportunities for future research.

The most commonly applied OR techniques are queueing theory and simulation. The strength of queueing models is that straightforward formulas provide quick insights in the trade-off between occupancy and blocking probability, delay, or overflow. Simulation models can incorporate more details, but require more development time and the results are often difficult to generalize to other wards or hospitals. Using optimization models, like dynamic or mathematical programming, to analyze and optimize hospital wards seems a promising direction for future research, as literature in this direction is relatively sparse.

When it comes down to integrating the OR models into practice there is little research available. Also, the literature reviewed in this chapter does not provide much insight to what extent these models are actually implemented and/or still used in practice. This may be explained by the fact that implementation requires different competences and techniques than solely OR. We are convinced that this final and for practitioners most important phase of an OR project should receive more attention both during OR projects and in OR literature. From our own experience we know it can be challenging to make the transition from model to practice, just as doing so the other way around. Focusing on factors for successful implementation we composed the following, non-inexhaustible, list:

- Stakeholders perceive a problem
- Stakeholders are willing to and prepared for change
- The chain of command is involved
- Stakeholders are involved with every phase of the analysis
- The team defines a clear set of key performance indicators

– The team thoroughly executes data collection, model verification and validation
– The team explains practical implications of model to stakeholders
– The team takes pre and post outcome samples on the key performance indicators in order to objectively compare the effects of the implemented model in practice.

Based on the number of references per type of ward, it is also clear where the opportunities lie for OR research on wards: AMUs, OBS, and WDWs. We are confidently optimistic that this contribution guides both researchers and health care professionals through the possibilities and opportunities OR offers for wards taking trade-offs between outcomes into account.

Acknowledgements The authors would like to express their gratitude for the valuable comments of Job Kievit and Wilbert van den Hout. Additionally, we thank Aleida Braaksma for her input for one of the case studies in this project.

Appendix: OR Model Types

As background we introduce here the commonly used OR models for analyzing the performance of hospital wards. The model categories are based on the ones applied in the ORchestra database (Hulshof et al. 2011). The ORchestra database, which distinguishes the following categories: algorithms, mathematical programming, dynamic programming, regression, time series, Markov models, stochastic models, queueing theory, and simulation. We define each of the OR techniques on a basic level, and provide introductory examples.

In this section we describe the more commonly used OR models in the context of a hospital ward setting: whereas OR researchers address 'servers' we use the term 'beds', and the 'customers' are referred to as 'patients'.

Algorithms Any procedure that follows predefined steps may be called an algorithm. Algorithms are often used for solving optimization problems, and are either based upon an exact mathematical analysis, or upon some heuristic rationale. Exact algorithms return an optimal solution but have significant long runtime, while heuristics approximate the optimal solution to decrease the runtime.

Algorithms are often applied to scheduling problems. The most simple illustration of a scheduling heuristic is the 'greedy algorithm', which prescribes that we schedule every patient at the earliest available bed or appointment slot. The 'earliest due date first' heuristic schedules the patients from the waiting list at the first available resource according to ascending maximum access times. Exact algorithms are typically more complicated than heuristics, so heuristics are often preferred for practical implementations. For more information on scheduling algorithms, the reader is referred to the book of Pinedo (2015).

Mathematical Programming Mathematical programming is the name given to a variety of related fields with a common form: the optimization of one or more

objectives subject to a set of limitations, called constraints. These fields include (non-)linear (integer) programming, stochastic programming, and network flow problems. The most commonly used of these is the field of linear programming, in which the objective function and the constraints are all linear functions of the decision variables, which can be stated as follows. One seeks to optimize (that is, maximize or minimize) a single objective, which is a linear function of a vector x of decision variables (that is, variables whose values we have some control over). The solution space of x is subject to a series of linear constraints, which state the operational limitations under which the system must operate. In matrix form, a linear program to maximize the objective can be stated as:

$$\begin{aligned} &\max z = cx \\ &\text{subject to}: \quad Ax \geq b \\ &\qquad\qquad\quad x \geq 0. \end{aligned}$$

Here, c is a row vector containing the reward rates per unit increase in a particular decision variable, A is the matrix whose rows contain the coefficients for the decision variables in the various constraints, and b is the column vector of right hand sides representing the limits for these various constraints.

A more practical example of this model is given in Sect. 5.4.4. For more information, see Winston (2003). A related yet distinct area frequently used in health care applications is the field of dynamic programming, which we consider next.

Dynamic Programming All sequential decision making problems are aggregated in the dynamic programming category. This type of models break the overall decision problem into a series of more easily solved sequential problems, consisting of the different phases at which a decision maker should choose one of the available actions. In each phase the 'system' under consideration is in a certain state, where the state contains enough information to decide which action would result in the best possible outcome for the system. The chosen action may result in direct costs, and determines the state of the system in the next phase, either with certainty or known likelihood. This can be stated more formally as follows: denote the phases by t, the states by i, the possible actions by a, the direct costs associated with action a when in state i by $c(i,a)$, the probability to go from state i to j when action a is chosen by $p(i,j|a)$, and the value function $V_n(i)$. A dynamic programming model may minimize costs, or maximize rewards. A dynamic programming model (here stated with the first objective) is optimized backwards by the recursion:

$$V_n(i) = \min_a \left\{ c(i,a) + \sum_j p(j,i|a) V_{n+1}(j) \right\}.$$

Markov decision models are related to dynamic programming models. However, whereas dynamic programming works backwards in time (from phase $n+1$ to n), Markov decision problems are solved forwards in time (from phase n to $n+1$).

Dynamic programming models are therefore more suitable for problems with a given deadline, where Markov decision theory is often applied to problems with infinite horizon. For more information, see Winston (2003).

Regression and Time Series Forecasting methods are used to forecast future values of a certain variable (or variables) based on historical data. Time series models such as 'moving average' and 'exponential smoothing' take a certain number of measurements as input for the forecast. Suppose we want to estimate x_t, the average occupancy of a ward on day t. We have data on the average occupancy for each day $1, 2, \ldots, t-1$. The moving average model and exponential smoothing models are given by:

$$x_t = \frac{\sum_{i=t-N}^{t-1} x_i}{N}, \qquad A_t = \alpha x_t + (1-\alpha)A_{t-1}$$

with N the number of days used to calculate the moving average, to be determined by the user, A_t be the smoothed average of the average occupancy at day t, with $A_0 = x_0$ as starting value, and $0 < \alpha < 1$ the smoothing factor.

Regression analysis estimates the relationship between the dependent variable that we wish to forecast, x_t, and (multiple) independent variables (y_t). The linear regression model is the most simple, and is described by:

$$x_t = \beta_0 + \beta_1 y_t + \epsilon_t.$$

Here, β_0, β_1 are coefficients that set the relationship between x and y, and ϵ_t is an error term. The coefficients β_0 and β_1 should be estimated to best fit the historic occupancy, and may be determined through the least squares method.

Statistics packages such as SPSS® and Minitab® contain most forecasting tools, and also Microsoft Excel® contains formulas for forecasting. For more information, see Winston (2003) and Hamilton (1994).

Markov and Stochastic Models A stochastic model is a description of the relation between random variables, whose values are not known with certainty beforehand. A random variable measured at discrete time points, e.g. each day at 10:00, is called a discrete-time random variable. A continuous time random variable is measured continuously, for example a patient's heart rate or temperature.

Markov models are a specific type of stochastic model, and have the property that the next value in the stochastic process is independent of its past, given its current value. An example of a Markov model is the outcome of a coin toss. We use the term 'stochastic model' for all stochastic models that do not have this property and do not fall into one of the other model categories. For more information, see Ross (2007) and Winston (2003).

Queueing Theory Queueing theory is the study of waiting lines in production systems. These systems consist of a waiting line and one or multiple servers, and are defined by an arrival and service process, see Fig. 5.5.

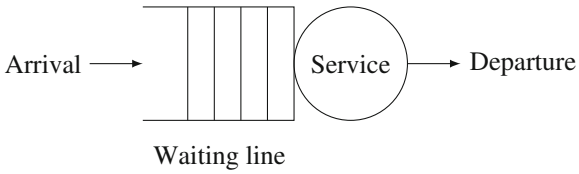

Fig. 5.5 A simple queue

A short way of referring to queues is by Kendall's notation: $A/B/s(/c)$, where A and B denote the arrival and service process, respectively, s is the number of servers, and c is an optional argument that denotes the number of places in the waiting line if this number is limited. Most queueing models assume Poisson arrivals, for which $A = M$. The service time distribution may be deterministic (D), exponential (E) or general (G). Typical performance measures that may be evaluated using queueing models are blocking probabilities, occupancy, throughput, patient waiting times, and bed idle times. Section 5.4 contains several examples of queueing models.

The QTS tool developed by Gross et al. (2008) is convenient for obtaining performance measures for most queueing (network) models with homogeneous arrival and service rates. For additional basic information on the queueing models described in this section, see Zonderland and Boucherie (2012) and Winston (2003).

Simulation Simulation models are used to mimic the evolution of a system over time, and consist of a list of what-if rules and procedures. We distinguish among discrete event simulation, Monte Carlo simulation, and system dynamics models. Discrete event simulations are event-driven routines, in which an event list is kept that contains the time stamps and types of events that will occur on those time stamps. With Monte Carlo simulation, repeated sampling from a probability distribution is carried out to obtain information on relevant performance measures. System dynamics models focus on the way different entities of the model influence each other, which relations are captured in a system of coupled, often non-linear differential equations.

Different simulation software packages exist, with different requirements regarding the user's programming abilities. Graphical simulation tools can often support the model validation process as the practitioners can see how the patients for example walk through the clinic. A drawback of graphical simulation models is that computation speed is reduced compared to non-graphical simulation packages.

For more information on simulation models, see Law (2007) and Winston (2003).

References

Akcali E, Murray JC, Lin C (2006) A network flow approach to optimizing hospital bed capacity decisions. Health Care Manag Sci 9(4):391–404

Akkerman R, Knip M (2004) Reallocation of beds to reduce waiting time for cardiac surgery. Health Care Manag Sci 7(2):119–126

Bagust A, Place M, Posnett JW (1999) Dynamics of bed use in accommodating emergency admissions: stochastic simulation model. BMJ 319(7203):155–158
Bailey NTJ (1952) A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. J R Stat Soc Ser B (Methodological) 14(2):185–199
Bekker R, de Bruin AM (2010) Time-dependent analysis for refused admissions in clinical wards. Ann Oper Res 178(1):45–65
Bekker R, Koeleman PM (2011) Scheduling admissions and reducing variability in bed demand. Health Care Manag Sci 14(3):237–249
Best TJ, Sandikci B, Eisenstein D, Meltzer D (2015) Managing hospital bed capacity through partitioning care into focused wings. Manuf Serv Oper Manag 17(2):157–176
Blair EL, Lawrence CE (1981) A queueing network approach to health care planning with an application to burn care in New York state. Socio-Econ Plan Sci 15(5):207–216
Bountourelis T, Luangkesorn L, Schaefer A, Maillart L, Nabors SG, Clermont G (2011) Development and validation of a large scale ICU simulation model with blocking. In: Himmelspach J, White KP, Fu M, Jain S, Creasey RR (eds) Proceedings of the 2011 winter simulation conference (WSC). Intensive Care Units. IEEE, pp 1143–1153
Bountourelis T, Ulukus MY, Kharoufeh JP, Nabors SG (2012) The modeling, analysis, and management of intensive care units. International series in operations research and management science, vol 184, book section 6. Springer, New York, pp 153–182
Braaksma A, Deglise-Hawkinson J, Denton BT, Van Oyen MP, Boucherie RJ, Mes MRK (2015) Online appointment scheduling with different urgencies and appointment lengths. In: Obtained through personal communication
Broyles JR, Cochran JK, Montgomery DC (2010) A statistical Markov chain approximation of transient hospital inpatient inventory. Eur J Oper Res 207(3):1645–1657
Chan CW, Farias VF, Bambos N, Escobar GJ (2012) Optimizing intensive care unit discharge decisions with patient readmissions. Oper Res 60(6):1323–1341
Chernew ME, Newhouse JP (2012) Health care spending growth. Handb Health Econ 2:1–43
Cochran JK, Bharti A (2006) Stochastic bed balancing of an obstetrics hospital. Health Care Manag Sci 9(1):31–45
Conforti D, Guerriero F, Guido R, Cerinic MM, Conforti ML (2011) An optimal decision making model for supporting week hospital management. Health Care Manag Sci 14(1):74–88
Cooke MW, Higgins J, Kidd P (2003) Use of emergency observation and assessment wards: a systematic literature review. Emerg Med J 20(2):138–142
Costa AX, Ridley SA, Shahani AK, Harper PR, De Senna V, Nielsen MS (2003) Mathematical modelling and simulation for planning critical care capacity*. Anaesthesia 58(4):320–327
Dantzig GB (1963) Linear programming and extensions. Princeton university press, Princeton, NJ
Davies R (1994) Simulation for planning services for patients with coronary artery disease. Eur J Oper Res 72(2):323–332
de Bruin AM, van Rossum AC, Visser MCC, Koole GM (2007) Modeling the emergency cardiac in-patient flow: an application of queuing theory. Health Care Manag Sci 10(2):125–137
De Bruin AM, Bekker R, van Zanten L, Koole GM (2010) Dimensioning hospital wards using the Erlang loss model. Ann Oper Res 178(1):23–43
Dobson G, Lee H, Pinker E (2010) A model of ICU bumping. Oper Res 58(6):1564–1576
Dumas MB (1985) Hospital bed utilization: an implemented simulation approach to adjusting and maintaining appropriate levels. Health Serv Res 20(1):43–61
El-Darzi E, Vasilakis C, Chaussalet T, Millard PH (1998) A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. Health Care Manag Sci 1(2):143–149
Ferreira RB, Coelli FC, Pereira WC, Almeida RMVR (2008) Optimizing patient flow in a large hospital surgical centre by means of discrete-event computer simulation models. J Eval Clin Pract 14(6):1031–1037
Gallivan S, Utley M (2011) A technical note concerning emergency bed demand. Health Care Manag Sci 14(3):250–252

Garg L, McClean S, Barton M, Meenan B, Fullerton K (2010) Forecasting hospital bed requirements and cost of care using phase type survival trees. In: 2010 5th IEEE international conference intelligent systems (IS). Bed Occupancy, Cardiology, pp 185–190

Garrison GM, Pecina JL (2016) Using the M/G/∞ queueing model to predict inpatient family medicine service census and resident workload. Health Informatics Journal 22(3):429–439

Gorunescu F, McClean SI, Millard PH (2002a) A queueing model for bed-occupancy management and planning of hospitals. J Oper Res Soc 53(1):19–24

Gorunescu F, McClean SI, Millard PH (2002b) Using a queueing model to help plan bed allocation in a department of geriatric medicine. Health Care Manag Sci 5(4):307–312

Gorunescu M, Gorunescu F, Prodan A (2002c) Continuous-time Markov model for geriatric patients behavior. Optimization of the bed occupancy and computer simulation. Korean J Comput Appl Math 9(1):185–195

Green L, Nguyen V (2001) Strategies for cutting hospital beds: the impact on patient service. Health Serv Res 36:421–442

Griffin J, Xia S, Peng S, Keskinocak P (2012) Improving patient flow in an obstetric unit. Health Care Manag Sci 15(1):1–14

Griffiths JD, Knight V, Komenda I (2013a) Bed management in a critical care unit. IMA J Manag Math 24(2):137–153

Griffiths JD, Williams JE, Wood RM (2013b) Modelling activities at a neurological rehabilitation unit. Eur J Oper Res 226(2):301–312

Gross D, Shortle JF, Thompson JM, Harris CM (2008) Qtsplus software, chapter Appendix E, 4th edn. Wiley, Hoboken, NJ, pp 489–492

Gunal MM, Pidd M (2010) Discrete event simulation for performance modelling in health care: a review of the literature. J Simul 4(1):42–51

Hamilton JD (1994) Time series analysis. Princeton University Press, Princeton, NJ

Harper PR, Pitt MA (2004) On the challenges of healthcare modelling and a proposed project life cycle for successful implementation. J Oper Res Soc 55(6):657–661

Harper PR, Shahani AK (2002) Modelling for the planning and management of bed capacities in hospitals. J Oper Res Soc 53(1):11–18

Harper PR, Knight VA, Marshall AH (2012) Discrete conditional phase-type models utilising classification trees: application to modelling health service capacities. Eur J Oper Res 219(3):522–530

Harris RA (1986) Hospital bed requirements planning. Eur J Oper Res 25(1):121–126

Harrison GW, Shafer A, Macky M (2005) Modelling variability in hospital bed occupancy. Health Care Manag Sci 8(4):325–334

Holm LB, Luras H, Dahl FA (2013) Improving hospital bed utilisation through simulation and optimisation: with application to a 40 Norwegian general hospital. Int J Med Inform 82(2):80–89

Hulshof PJH, Boucherie RJ, van Essen JT, Hans EW, Hurink JL, Kortbeek N, Litvak N, Vanberkel PT, van der Veen E, Veltman B, Vliegen IMH, Zonderland ME (2011) ORchestra: an online reference database of OR/MS literature in health care. Health Care Manag Sci 14(4):383–384

Isken MW, Ward TJ, Littig SJ (2011) An open source software project for obstetrical procedure scheduling and occupancy analysis. Health Care Manag Sci 14(1):56–73

Keepers K, Harrison GW (2009) Internal flows and frequency of internal overflows in a large teaching hospital. In: McClean S, Millard P, El-Darzi E, Nugent C (eds) Intelligent patient management. Studies in computational intelligence, vol 189. Springer, Berlin/Heidelberg

Kim SC, Horowitz I, Young KK, Buckley TA (1999) Analysis of capacity management of the intensive care unit in a hospital. Eur J Oper Res 115(1):36–46

Kim SC, Horowitz I, Young KK, Buckley TA (2000) Flexible bed allocation and performance in the intensive care unit. J Oper Manag 18(4):427–443

Kokangul A (2008) A combination of deterministic and stochastic approaches to optimize bed capacity in a hospital unit. Comput Methods Prog Biomed 90(1):56–65

Kolker A (2013) Interdependency of hospital departments and hospital-wide patient flows. International series in operations research and management science, book section 2, vol 206. Springer, Boston, pp 43–63

Kortbeek N, Braaksma A, Smeenk FHF, Bakker PJM, Boucherie RJ (2015) Integral resource capacity planning for inpatient care services based on bed census predictions by hour. J Oper Res Soc 66(7):1061–1076

Kumar S (2011) Modeling hospital surgical delivery process design using system simulation: optimizing patient flow and bed capacity as an illustration. Technol Health Care 19(1):1–20

Kumar A, Mo J (2010) 'Models for bed occupancy management of a hospital in Singapore. In: Ali A (ed) Proceedings of the 2010 International conference on industrial engineering and operations management, Dhaka, pp 1–6

Kusters RJ, Groot PMA Modelling resource availability in general hospitals design and implementation of a decision support model. Eur J Oper Res 88(3):428–445

Landa P, Sonnessa M, Tanfani E, Testi A (2014) A discrete event simulation model to support bed management. In: 2014 International conference on simulation and modeling methodologies, technologies and applications (SIMULTECH). Bed Occupancy, Emergency Service, Hospital, pp 901–912

Lapierre SD, Goldsman D, Cochran R, DuBow J (1999) Bed allocation techniques based on census data. Socio-Econ Plan Sci 33(1):25–38

Law AM (2007) Simulation modeling and analysis, 4th edn. McGraw-Hill, New York

Li X, Beullens P, Jones D, Tamiz M (2009) An integrated queuing and multi-objective bed allocation model with application to a hospital in China. J Oper Res Soc 60:330–338

Litvak N, van Rijsbergen M, Boucherie RJ, van Houdenhoven M (2008) Managing the overflow of intensive care patients. Eur J Oper Res 185(3):998–1010

Mackay M (2001) Practical experience with bed occupancy management and planning systems: an Australian view. Health Care Manag Sci 4(1):47–56

Mackay M, Lee M (2005) Choice of models for the analysis and forecasting of hospital beds. Health Care Manag Sci 8(3):221–230

Mallor F, Azcarate C (2014) Combining optimization with simulation to obtain credible models for intensive care units. Ann Oper Res 221(1):255–271

Mallor F, Azcárate C, Barado J (2016) Flex Serv Manuf J 28(1–2):62–89. https://doi.org/10.1007/s10696-014-9209-8

Marcon E, Kharraja S, Smolski N, Luquet B, Viale JP (2003) Determining the number of beds in the postanesthesia care unit: a computer simulation flow approach. Anesth Analg 96(5):1415–1423

Marmor YN, Rohleder TR, Cook DJ, Huschka TR, Thompson JE (2013) Recovery bed planning in cardiovascular surgery: a simulation case study. Health Care Manag Sci 16(4):314–327

Massey WA, Whitt W (1994) An analysis of the modified offered-load approximation for the nonstationary erlang loss model. Ann Appl Probab 4(4):1145–1160

Masterson BJ, Mihara TG, Miller G, Randolph SC, Forkner ME, Crouter AL (2004) Using models and data to support optimization of the military health system: a case study in an intensive care unit. Health Care Manag Sci 7(3):217–224

McClain JO (1978) A model for regional obstetric bed planning. Health Serv Res 13(4):378–394

McManus ML, Long MC, Cooper A, Litvak E (2004) Queuing theory accurately models the need for critical care resources. Anesthesiology 100(5):1271–1276

Mustafee N, Lyons T, Rees P, Davies L, Ramsey M, Williams MD (2012) Planning of bed capacities in specialized and integrated care units: incorporating bed blockers in a simulation of surgical throughput. In: Proceedings of the 2012 winter simulation conference (WSC). IEEE, Piscataway, NJ, pp 1–12

Nguyen JM, Six P, Parisot R, Antonioli D, Nicolas F, Lombrail P (2003) A universal method for determining intensive care unit bed requirements. Intensive Care Med 29(5):849–852

Oddoye JP, Yaghoobi MA, Tamiz M, Jones DF, Schmidt P (2007) A multi-objective model to determine efficient resource levels in a medical assessment unit. J Oper Res Soc 58(12):1563–1573

Oddoye JP, Jones DF, Tamiz M, Schmidt P (2009) Combining simulation and goal programming for healthcare planning in a medical assessment unit. Eur J Oper Res 193(1):250–261
OECD. Health expenditure and financing. http://stats.oecd.org/
Pehlivan C, Augusto V, Xiaolan X, Crenn-Hebert C (2012) Multi-period capacity planning for maternity facilities in a perinatal network: a queuing and optimization approach. In: 2012 IEEE international conference on automation science and engineering (CASE). Bed Occupancy, Obstetrics, pp 137–142
Pinedo ML (2015) Scheduling: theory, algorithms, and systems, 5th edn. Springer, New York
Ramakrishnan M, Sier D, Taylor PG (2005) A two-time-scale model for hospital patient flow. IMA J Manag Math 16(3):197–215
Ridge JC, Jones SK, Nielsen MS, Shahani AK (1998) Capacity planning for intensive care units. Eur J Oper Res 105(2):346–355
Roelofs M, Bisschop J (2012) AIMMS: the user's guide, chapter 20. Paragon Decision Technology, Haarlem, pp 235–244
Ross SM (2007) Introduction to probability models, 9th edn. Academic, San Diego, CA
Scott I, Vaughan L, Bell D (2009) Effectiveness of acute medical units in hospitals: a systematic review. Int J Qual Health Care 21(6):397–407
Shahani AK, Ridley SA, Nielsen MS (2008) Modelling patient flows as an aid to decision making for critical care capacities and organisation. Anaesthesia 63(10):1074–1080
Shmueli A, Sprung CL, Kaplan EH (2003) Optimizing admissions to an intensive care unit. Health Care Manag Sci 6(3):131–136
Shonick W, Jackson JR (1973) An improved stochastic model for occupancy-related random variables in general-acute hospitals. Oper Res 21(4):952–965
Sissouras AA, Moores B (1976) The optimum number of beds in a coronary care unit. Omega 4(1):59–65
Swain RW, Kilpatrick KE, Marsh JJ (1977) Implementation of a model for census prediction and control. Health Serv Res 12(4):380–395
Taylor GJ, McClean SI, Millard PH (2000) Stochastic models of geriatric patient bed occupancy behaviour. J R Stat Soc: Ser A Stat Soc 163(1):39–48
Troy PM, Rosenberg L (2009) Using simulation to determine the need for ICU beds for surgery patients. Surgery 146(4):608–617
Utley M, Gallivan S, Treasure T, Valencia O (2003a) Analytical methods for calculating the capacity required to operate an effective booked admissions policy for elective inpatient services. Health Care Manag Sci 6(2):97–104
Utley M, Gallivan S, Davis K, Daniel P, Reeves P, Worrall J (2003b) Estimating bed requirements for an intermediate care facility. Eur J Oper Res 150(1):92–100
Utley M, Gallivan S, Jit M (2005) How to take variability into account when planning the capacity for a new hospital unit. In: Health operations management: patient flow logistics in health care, pp 146–161
van Dijk NM, Kortbeek N (2009) Erlang loss bounds for OT-ICU systems. Queueing Syst 63(1–4):253–280
van Essen JT, van Houdenhoven M, Hurink JL (2015) Clustering clinical departments for wards to achieve a prespecified blocking probability. OR Spectrum 37(1):243–271
Vanberkel P, Blake J (2007) A comprehensive simulation for wait time reduction and capacity planning applied in general surgery. Health Care Manag Sci 10:373–385
Vanberkel PT, Boucherie RJ, Hans EW, Hurink JL, van Lent WAM, van Harten WH (2011) An exact approach for relating recovering surgical patient workload to the master surgical schedule. J Oper Res Soc 62(10):1851–1860
Vasilakis C, El-Darzi E (2001) A simulation study of the winter bed crisis. Health Care Manag Sci 4(1):31–36
Vasilakis C, El-Darzi E, Chountas P (2008) A decision support system for measuring and modelling the multi-phase nature of patient flow in hospitals. Studies in computational intelligence, book section 12, vol 109. Springer, Berlin/Heidelberg, pp 201–217

Vissers J, Beech R (2005) Health operations management: patient flow logistics in health care. Routledge, New York

Williams J, Dumont S, Parry-Jones J, Komenda I, Griffiths J, Knight V (2015) 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

Winston WL (2003) Operations research: applications and algorithms, 4th edn. Brooks/Cole–Thomson Learning, Belmont, CA

Yang M, Fry MJ, Raikhelkar J, Chin C, Anyanwu A, Brand J, Scurlock C (2013) A model to create an efficient and equitable admission policy for patients arriving to the cardiothoracic ICU. Crit Care Med 41(2):414–422

Zhu Z (2011) Impact of different discharge patterns on bed occupancy rate and bed waiting time: a simulation approach. J Med Eng Technol 35(6–7):338–343

Zhu Z (2014) An online short-term bed occupancy rate prediction procedure based on discrete event simulation. J Hosp Adm 3(4):p37

Zonderland ME, Boucherie RJ, Carter MW, Stanford DA (2015) Modeling the effect of short stay units on patient admissions. Oper Res Health Care 5:21–27

Zonderland ME, Boucherie RJ (2012) Queuing networks in health care systems, book section 9. Springer, New York, pp 201–243