MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
*Articles in Advance*, pp. 1–16
ISSN 1523-4614 (print), ISSN 1526-5498 (online)
http://pubsonline.informs.org/journal/msom

# Bed Blocking in Hospitals Due to Scarce Capacity in Geriatric Institutions—Cost Minimization via Fluid Models

**Noa Zychlinski,[a] Avishai Mandelbaum,[a] Petar Momčilović,[b] Izack Cohen[a]**

[a] Faculty of Industrial Engineering and Management, Technion—Israel Institute of Technology, Haifa 32000, Israel; [b] Department of Industrial and Systems Engineering, Texas A&M University, College Station, Texas 77843

**Contact:** noazy@tx.technion.ac.il, http://orcid.org/0000-0002-5125-3089 (NZ); avim@ie.technion.ac.il (AM); petar@tamu.edu (PM); izik68@techunix.technion.ac.il (IC)

**Received:** February 9, 2017
**Revised:** March 8, 2018; July 19, 2018
**Accepted:** July 29, 2018
**Published Online in Articles in Advance:** May 21, 2019

**https://doi.org/10.1287/msom.2018.0745**



**Abstract.** ***Problem definition***: This research focuses on elderly patients who have been hospitalized and are ready to be discharged, but they must remain in the hospital until a bed in a geriatric institution becomes available; these patients "block" a hospital bed. Bed blocking has become a challenge to healthcare operators because of its economic implications and the quality-of-life effect on patients. Indeed, hospital-delayed patients who do not have access to the most appropriate treatments (e.g., rehabilitation) prevent new admissions. Moreover, bed blocking is costly, because a hospital bed is more expensive to operate than a geriatric bed. We are thus motivated to model and analyze the flow of patients between hospitals and geriatric institutions to improve their joint operation. ***Academic/practical relevance***: Practically, our joint modeling of hospital-institution is necessary to capture blocking effects. In contrast to previous research, we address an entire time-varying network, which enables an explicit consideration of blocking costs. Theoretically, our fluid model captures blocking without the need for reflection, which simplifies the analysis as well as the convergence proof of the corresponding stochastic model. ***Methodology***: We develop a mathematical fluid model, which accounts for blocking, mortality, and readmission—all significant features of the discussed environment. Then, for bed allocation decisions, the fluid model and especially, its offered load counterpart turn out insightful and easy to implement. ***Results***: The comparison between our fluid model, a two-year data set from a hospital chain, and simulation results shows that our model is accurate and useful. Moreover, our analysis yields a closed form expression for bed allocation decisions, which minimizes the sum of underage and overage costs. Solving for the optimal number of geriatric beds in our system shows that significant reductions in cost and waiting list length are achievable compared with current operations. ***Managerial implications***: Our model can support healthcare managers in allocating geriatric beds to reduce operational costs. Moreover, our model facilitates three extensions: a periodic reallocation of beds, incorporation of setup costs into bed allocation decisions, and accommodating home care (or virtual hospitals) when feasible.

**Funding:** The work of N. Zychlinski has been partially supported by the Israeli Ministry of Science, Technology and Space and Technion—Israel Institute of Technology. The work of A. Mandelbaum has been partially supported by the Israel Science Foundation [Grants 357/80 and 1955/15]. The work of A. Mandelbaum and P. Momčilović has been partially supported by the United States–Israel Binational Science Foundation [Grant 2014180]. The work of P. Momčilović has been partially supported by the NSF Division of Civil, Mechanical and Manufacturing Innovation [Grant 1362630].
**Supplemental Material:** The online appendices are available at https://doi.org/10.1287/msom.2018.0745.

**Keywords:** bed blocking • bed planning for long-term care facilities • geriatric institutions • fluid models • time-varying queueing networks with blocking

## 1. Introduction

Providing high-quality healthcare services for the aging population is becoming a major challenge in developed countries. This challenge is amplified by the fact that the number of elderly people ages 65 years old and over, who today account for 10% of the population, will double within two decades (United Nations Population Fund 2014, World Health Organization 2014). Moreover, elderly patients are often frail and undergo frequent hospitalizations. These facts are and will increasingly be major contributors to the high occupancy levels in inpatient wards and emergency departments (EDs). For example, in the last several years, some Organization for Economic Co-operation and Development (OECD) countries reported averages of over 90% occupancy levels in hospital inpatient wards (OECD iLibrary 2013, NHS England 2015), and these yearly averages hardly reveal the hour-by-hour reality of the busiest periods (e.g., winters).

The bed blocking problem occurs when hospital patients are ready to be discharged but must remain in the hospital until a bed in a more appropriate geriatric facility (a nursing home or a geriatric institution) becomes available. Research about the bed blocking problem (e.g., Rubin and Davies 1975, Namdaran et al. 1992, El-Darzi et al. 1998, Koizumi et al. 2005, Cochran and Bharti 2006, Travers et al. 2008, Osorio and Bierlaire 2009, Shi et al. 2015) is important, because it can potentially improve the quality of patient care and reduce the mounting costs associated with bed blocking (Cochran and Bharti 2006). For example, the estimated cost of bed blocking in the United Kingdom alone exceeds $1.2 billion per year (BBC News 2016). In this paper, we focus on the bed blocking problem caused by bed shortage in geriatric institutions rather than in general nursing homes, because in our setting and according to the data that we analyze, the problem in geriatric institutions is much more severe. Having said that, our modeling framework accommodates any environment in which the phenomenon of blocking is severe and gives rise to operational challenges.

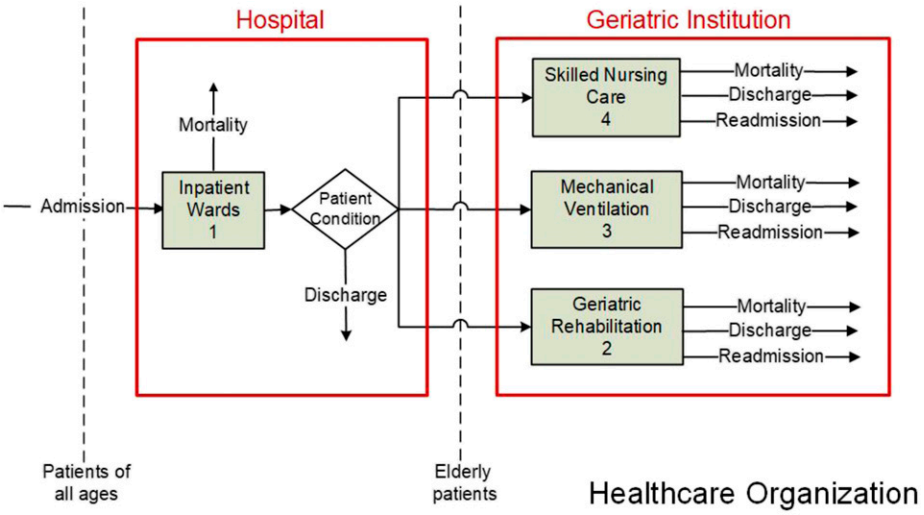
In contrast to previous models, which relied on simulations for modeling bed blocking, our research offers an analytical model for minimizing the overage and underage costs of a system consisting of hospitals and geriatric institutions; the model yields a tractable solution by determining the optimal number of beds for each geriatric ward.

We focus on long-term geriatric bed allocation by considering the environment described in Figure 1: it covers inpatient wards in hospitals and geriatric institutions. In our setting, the central decision maker is a large healthcare organization, which operates several hospitals and several geriatric institutions. In some countries (e.g., Singapore and Israel), the government functions as this organization. In England, the National Health Service (NHS), an arm of the government, is the central decision maker; in Australia, it is the Medicare Healthcare System, and in the United States, it can be the Veterans Administration with its 500+ hospitals.

Congestion problems and their highly significant effect, both medically and financially, motivated us to model and analyze the system, which is depicted schematically in Figure 1. Patient flow begins when people of all ages are admitted to hospital inpatient wards (Station 1). On treatment completion and focusing on geriatric patients, hospital doctors decide whether the patient is capable of returning to the community or requires additional care in a geriatric institution. We subdivide the latter option into the three most common long-term care geriatric wards: rehabilitation (Station 2), mechanical ventilation (Station 3), and skilled nursing care (Station 4).

Patients who are sent to a geriatric rehabilitation ward stay there for one month on average before they are able to return to full or partial functioning. Mechanical ventilation wards treat patients who cannot breathe on their own, typically after three unsuccessful weaning attempts in a hospital; the average stay in a mechanical ventilation ward is five to six months. Unfortunately, only a minority of these patients are discharged; most die or are readmitted to hospitals. Skilled nursing wards treat patients who, in addition to functional dependency, suffer from active diseases that require close medical supervision (for example, because of bedsores or chemotherapy); the average stay there is 1–1.5 months. Some patients are discharged to nursing homes, but again, most either die or are readmitted to hospitals. According to our data and the

**Figure 1.** (Color online) Network of Patient Flow Through Inpatient Wards and Geriatric Institutions




*Note*. The readmission arrows substitute for arrows from Station 2, 3, or 4 back to Station 1.

doctors and managers with whom we consulted, transfers between geriatric wards of different types rarely happen.

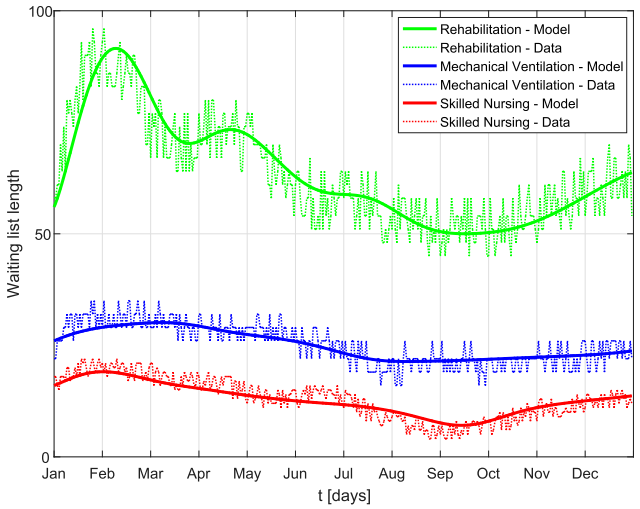
The methodology that we propose is rather general and can accommodate other settings with different numbers or types of wards. Because the system that we analyze and the data that we use are for three types of geriatric wards, in the empirical part of the paper, we focus on the four stations depicted in Figure 1. Applying our general methodology to analyzing these stations, for which there are long waiting lists, will yield policies that significantly reduce total operational costs. (Figure 2 in Online Appendix B is a schematic analog of Figure 1.)

To this end, we develop a mathematical fluid model that accounts for blocking, mortality, and readmission—all significant features of the discussed environment. Then, we use our fluid model and its time-varying offered load counterpart to formulate and solve bed allocation problems for geriatric wards. Our goal is to find the optimal number of geriatric beds to minimize the total overage plus underage costs of the system. Moreover, we propose two feasible extensions for capacity allocation problems with time-varying demand of beds: a periodic reallocation of beds and the incorporation of setup costs into bed allocation decisions.

In our analysis, we use two data sets over a period of two years. The first covers the patient flow in a hospital chain made up of four hospitals and three geriatric institutions (three rehabilitation wards, two mechanical ventilation wards, and three skilled nursing wards). The second data set includes individual in-hospital waiting lists for each geriatric ward. (Details about our data are provided in Online Appendix A.) These data indicate that the average in-hospital waiting times are 28 days for mechanical ventilation, 17 days for skilled nursing care, and 3.5 days for rehabilitation wards. Although the average waiting time for rehabilitation seems relatively short, this is definitely not the case when considering the fact that these are elderly patients waiting unnecessarily for their rehabilitation care while occupying a bed that could have been used for newly admitted acute patients. Moreover, the numbers of patients who are referred to a rehabilitation ward are five and nine times those of the corresponding numbers for skilled nursing care and mechanical ventilation, respectively; this implies (Section 4.1) that the overall demand that they generate exceeds that of the other patients.

Figure 2 presents the waiting list length (daily resolution) within the hospital for each geriatric ward over one calendar year. The dotted lines in Figure 2 represent length according to our data, whereas the solid lines in Figure 2 represent our fluid model (Equations (5) and (6) in Section 3.2). According to this plot, all three geriatric wards work at full capacity throughout the year (long waiting lists); furthermore, in the winter, the demand for beds increases.

**Figure 2.** (Color online) Waiting List Length in Hospital for Each Geriatric Ward—Model (Solid Lines) vs. Data (Dashed Lines)




*Notes.* The $x$ axis is one calendar year in units of days. (We are plotting here the second year of our data. The first year was used to fit the parameters of our model.)

The fit between our model and the data is excellent. In fact, in Online Appendix A, we show via multiple scenarios with various treatment distributions that our continuous, deterministic fluid model approximates well and usefully its underlying stochastic environment.

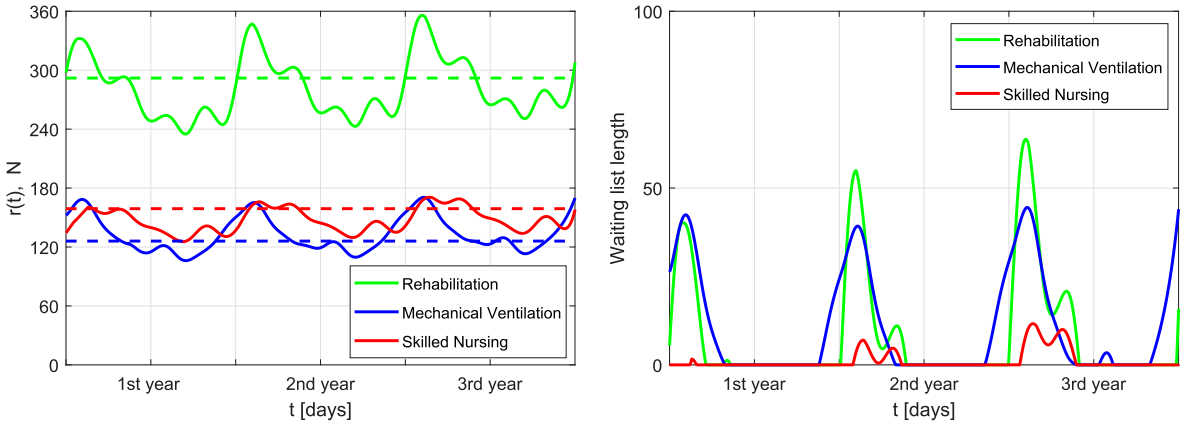
The long waiting lists and the fact that hospitalization costs are much higher in hospitals than in geriatric institutions indicate that the system is operated inefficiently; this leads to excessive costs that can be reduced by adopting our solution. Moreover, in Sections 5.1 and 6.2.1, we show how the constant and periodic allocations that we suggest can reduce costs and shorten waiting lists. (The latter is illustrated in Figure 3, right panel and Figure 4, lower left panel; this is relative to the current length of the waiting lists presented in Figure 2.)

### 1.1. Contributions

The main contributions of our research are as follows.

1. Modeling. We develop and analyze an analytical model comprising both long-term care geriatric wards and their feeding hospitals. This joint modeling is necessary to capture blocking effects (whereas previous research was restricted to a single-station utility maximization) (e.g., Jennings et al. 1997). We do so by explicitly considering geriatric ward blocking costs and minimizing the overall underage and overage costs within the system. Our approach has significant modeling strength beyond its base case. For example, it also accommodates periodic bed allocations and home care (virtual hospital) alternatives to mention just two relevant and important examples.

**Figure 3.** (Color online) Optimal Solution




*Notes.* In the left panel, the solid lines represent the offered load for each geriatric ward, and the dashed lines represent the optimal number of beds. In the right panel, the waiting list lengths in the hospital, according to the optimal solution, are depicted; this is relative to the current waiting list lengths presented in Figure 2.

2. Methodology. Our work contributes to the literature on queueing (fluid) networks with blocking. Indeed, as far as we know, this is the first paper to model, analyze, rigorously justify, and validate time-varying fluid network models with blocking. In particular, our proposed fluid model of a network captures blocking without the need for reflection (Section 2.2); it applies to general networks (for example, tandem networks), which simplifies theory (proofs and convergence) and numerical applications. Moreover, we use our model to derive analytical solutions and insights about cost minimization and bed allocation policies.

3. Practice. This research provides practical tools for bed planning of time-varying healthcare networks with blocking. Moreover, this research gives rise to novel capacity allocation strategies; it also quantifies the impact of emerging alternatives for care. Specifically, as already mentioned, we offer closed form solutions for periodic reallocation of beds that respond to seasonal demand, for diverting patients to home care, and for incorporating setup costs. These are but three examples made analyzable by our model.

4. Managerial insights. Our framework amplifies the need for an integrated view of patient flow within and beyond hospitals. This view is required to capture the costs of blocking, which are dramatically escalating with population aging. Our models yield managerial recommendations for healthcare managers in allocating geriatric beds and rectifying the bed blocking problem. The recommendations that we provide can also be implemented gradually (e.g., subject to budget constraints) while estimating the cost reduction at each step.

## 2. Literature Review

The review covers the main areas that are relevant to this research: high-level modeling of healthcare systems, queueing networks with blocking, time-varying queueing networks, and bed planning in long-term care facilities.

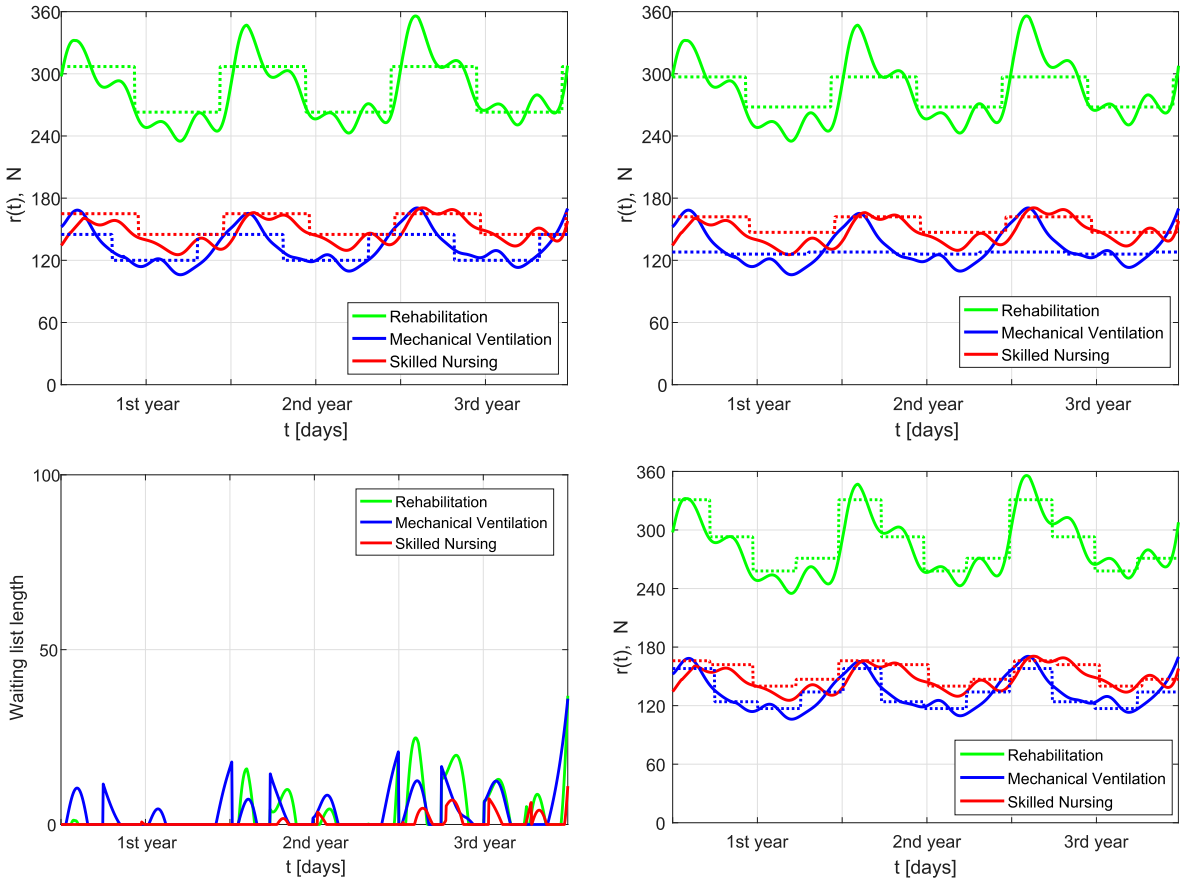
### 2.1. High-level Modeling of Healthcare Systems

The three main approaches used for modeling healthcare systems with elderly patients have been Markov models, system dynamics, and discrete event simulation.

For tractability reasons, Markov models have been applied to networks with a limited number of stations, typically two to three, to characterize steady-state performance, such as the length of stay (LOS) at each station. For example, Harrison and Millard (1991) analyze the empirical distribution of patient LOS in geriatric wards by fitting a sum of two exponentials to a data set: most patients are discharged or die shortly after admission, whereas some stay hospitalized for months. Several papers use Markov models to describe the flow of geriatric patients between hospitals and community-based care (Taylor et al. 1997, 2000; Faddy and McClean 2005; Xie et al. 2005; McClean and Millard 2006). In general, these models, which include short-stay and long-stay states in each facility, distinguish between the movement of patients within and between facilities. Unlike these papers, our approach emphasizes station capacity and time-varying parameters.

Another common approach for modeling healthcare systems is system dynamics. It is used to analyze patient flow through healthcare services by focusing on the need to coordinate capacity levels across all health services. Wolstenholme (1999) develops a patient flow model for the UK National Health Service and uses it to analyze alternatives for shortening waiting times of community care patients. According to the author, reducing total waiting times can be achieved by adding "intermediate care" facilities, which are aimed at preventing elderly medical patients from hospitalization and community care. Our approach contributes to this

**Figure 4.** (Color online) Optimal Reallocation of Beds When No Reallocation Costs Are Introduced (Upper Left Panel), When Reallocation Costs Are Introduced (Upper Right Panel), and When Four Reallocation Points Are Allowed (Lower Right Panel)




*Note.* Waiting list length under the optimal reallocation policy when no reallocation costs are introduced (lower left plot).

line of research by considering the dependency between capacity allocation and waiting time.

System dynamics is also used to analyze the bed blocking problem (Gray et al. 2006, Travers et al. 2008, Rohleder et al. 2013). These papers show the importance of coordinating capacity levels across different health services. Desai et al. (2008) use system dynamics to forecast the future demand for social care services by elderly people. Although our proposed fluid model is also deterministic, we are able to justify it as the fluid limit of an underlying stochastic model/system.

Discrete event simulation is another popular approach for analyzing complex systems and phenomena, such as bed blocking. El-Darzi et al. (1998) examine the impact of bed blocking and occupancy on patient flow through geriatric wards. They show that the availability of acute beds is strongly connected to referral rates for long-stay care facilities. Katsaliaki et al. (2005) build a simulation model of elderly patient flow between the community, hospitals, and geriatric institutions. They approximate the delay in discharge from hospital and the relevant costs. Armony et al. (2015) and Shi et al. (2015) discuss a two-timescale (days and hours) service time in hospital wards. Shi et al. (2015) investigate ED boarding times (waiting for admission to hospital wards) at a Singaporean hospital. Via simulation studies, they examine the effects of various discharge policies on admission waiting times. The two-timescale service time captures both treatment time and additional service time caused by operational factors, such as discharge schedule. In our research, we develop a time-varying analytical model for setting bed capacities in geriatric institutions. Our model evolves on a single timescale—days; because of the decisions in which we are interested (and the data that we have), days are natural and adequate.

## 2.2. Queueing Networks with Blocking

Several blocking mechanisms are acknowledged in the literature (Perros 1994, Balsamo et al. 2001). We focus on the *blocking after service* mechanism, which happens when a patient attempts to enter a fully capacitated Station $j$ on completion of treatment at Station $i$. Because it is not possible to queue in front of Station $j$, the patient must wait in Station $i$ and therefore, blocks a bed there until a departure occurs at Station $j$.

Healthcare systems usually have complex network topologies, multiple-server queues, and time-varying dynamics. In contrast, closed form solutions of queueing

models with blocking exist only for steady-state, single-server networks with two or three tandem queues or two cyclic queues (Osorio and Bierlaire 2009). The solutions for more complex networks are based on approximations, which are typically derived via decomposition methods (Hillier and Boling 1967, Takahashi et al. 1980, Gershwin 1987, Koizumi et al. 2005, Osorio and Bierlaire 2009) and expansion methods (Kerbache and MacGregor Smith 1987, 1988; Cheah and Smith 1994). Koizumi et al. (2005) use a decomposition method to analyze a healthcare system with mentally disabled patients as a multiple-server queueing network with blocking, whereas Osorio and Bierlaire (2009) develop an analytic finite capacity queueing network that enables the analysis of patient flow and bed blocking in a network of hospital operative and postoperative units.

Bretthauer et al. (2011) offer a heuristic method for estimating the waiting time for each station in a tandem queueing network with blocking by adjusting the per server service rate to account for blocking effects. Bekker and de Bruin (2010) analyze the effect of a predictable patient arrival pattern on a clinical ward regarding its performance and bed capacity requirements. In particular, the authors use the offered load approximation and the square root staffing formula for calculating the required beds for each day of the week. Although we also use the offered load approximation for the time-varying demand, our approach is different, because it goes beyond a single-station analysis and takes into account blocking effects by minimizing overage and underage costs. Moreover, the periodic reallocation that we suggest takes into account a reallocation cost that is associated with adding and removing a bed.

Capturing blocking in stochastic systems with a single station in steady state has been done via reflection. Specifically, reflection is a mathematical mechanism that has been found necessary to capture customer loss (Garnett et al. 2002; Whitt 2002, chapter 5.2). Reflection modeling, however, requires the use of indicators, which cause technical continuity problems when calculating approximating limits. We circumvent this challenge by developing a fluid model with blocking but without reflection, which enables us to prove convergence of our stochastic model without reflection. Our simple and intuitive model, compared with models with reflection, enables us to model, successfully and insightfully, time-varying networks.

### 2.3. Queueing Networks with Time-Varying Parameters

Time-varying queueing networks have been analyzed by McCalla and Whitt (2002), who focused on long service lifetimes, measured in years, in private-line telecommunication services. Liu and Whitt (2011b) analyze time-varying networks with many-server fluid queues and customer abandonment. In addition, time-varying queueing models have been analyzed for setting staffing requirements in service systems with unlimited queue capacity by using the offered load analysis (Whitt 2013). The methods for coping with time-varying demand when setting staffing levels are reviewed in Green et al. (2007) and Whitt (2007). A recent work of Li et al. (2015) focuses on stabilizing blocking probabilities in loss models with a time-varying Poisson arrival process by using a variant of the modified offered load approximation.

Fluid frameworks are well adapted to large time-varying overloaded systems (Mandelbaum et al. 1998, 1999), which is the case here. Previous research shows that fluid models have been successfully implemented in modeling healthcare systems (Ata et al. 2013, Cohen et al. 2014, Yom-Tov and Mandelbaum 2014). Moreover, fluid models yield analytical insights, which typically cannot be obtained using their alternatives (e.g., simulation, time-varying stochastic queueing networks).

### 2.4. Bed Planning for Long-term Care Facilities

Most research on bed planning in healthcare systems focuses on short-term facilities, such as hospitals (Green 2004, Akcali et al. 2006). Research about bed planning for long-term care facilities is scarce. We now review the existing literature.

Future demand for long-term care has a strong impact on capacity-setting decisions. Hare et al. (2009) develop a deterministic model for predicting future long-term care needs in home and community care services in Canada. Zhang et al. (2012) develop a simulation-based approach to find the minimal number of nursing home beds to achieve a target waiting time. The model that we suggest considers time-varying demand for beds throughout the year as well as mortality and readmission rates, which are all significant in the context of geriatric patients. In addition, we analyze a network capacity problem of several geriatric wards by taking into account blocking effects in hospitals.

De Vries and Beekman (1998) present a deterministic dynamic model for expressing waiting lists and waiting times of psychogeriatric patients for nursing homes based on data from the previous year. Ata et al. (2013) analyze the expected profit of hospice care. They propose an alternative reimbursement policy for the US Medicare and determine the recruiting rates of short- and long-stay patients to maximize profitability of the hospice. Kao and Tung (1981) consider the monthly fluctuation in demand for hospital services, but the bed allocation that they allow is constant throughout the year. In particular, they try to minimize the hospital yearly average overflow probability. To accommodate for the seasonal demand, we suggest a periodic reallocation of beds, taking into account the reallocation cost that is associated with adding and removing each bed.

Harrison and Zeevi (2005) develop a method, which was extended in Bassamboo et al. (2006), for staffing large call centers with multiple customer classes and multiple-server pools; they deploy stochastic fluid models to minimize the sum of personnel costs and abandonment penalties. The method that they suggest reduces the staffing problem to a multidimensional Newsvendor problem, and hence, the critical fractile solution that they suggest is distribution dependent. In Remark 3, we further elaborate on the relation of Harrison and Zeevi (2005) to this work.

Afèche et al. (2017) develop a fluid model for maximizing the profit of service firms by determining customer acquisition investment as well as a bottleneck capacity allocation among heterogeneous customer classes. This allocation and resulting service access quality affect customers' routing in the network. Our research includes finite capacities at all stations and time variation. This allows us to consider the blocking customers occupying servers in the first station and explicitly accommodate the blocking costs when calculating the optimal number of beds. Moreover, we justify the fluid model by proving convergence of the corresponding stochastic model.

Other related research is from the telecommunication field. Jennings et al. (1997) find the optimal number of leased private lines for profit maximization. Because of very long service times (years), their analysis is transient—the system does not reach steady state within the observation period. Because hospitalization time is long compared with the planning horizon, transient analysis is also relevant in the context of geriatric hospitalization.

## 3. The Model

In this section, we describe our environment and its dynamics. We then formally introduce model notation and equations.

### 3.1. Environment, Dynamics, and Notation

Consider the four stations in Figure 1: hospital wards (Station 1) and long-term care geriatric wards—rehabilitation (Station 2), mechanical ventilation (Station 3), and skilled nursing care (Station 4). Station 1 includes all ward patients, whereas Stations 2–4 include only geriatric patients who need long-term care beyond hospitalization.

Our model is at the macrolevel; thus, the capacity of each station is an aggregation of the individual capacities of all stations of this type in the discussed geographical area (e.g., assume that a district includes three rehabilitation wards; then, the capacity of the modeled rehabilitation station is the sum of all three individual capacities). Such aggregated capacities are justified, because in practice, patients can be sent from any individual hospital to any individual geriatric ward and vice versa, especially if they are all within the same geographic area (a city or a district).

Online Appendix I summarizes the notation that we use. We model the exogenous arrival rate to hospital wards as a continuous time-varying function $\lambda(t)$ (Mandelbaum et al. 1999). Internal arrivals are patients returning from geriatric wards back to the hospital. Hospital wards include $N_1$ beds. If there are available beds, arriving patients are admitted and hospitalized; otherwise, they wait in the queue. We assume that hospital wards have ample waiting rooms, because the ED serves as a buffer for them; nevertheless, our model can accommodate blocking of the first station (e.g., when ambulance diversion is significant enough). Patients leave the queue either when a bed becomes available or if they, unfortunately, die. Medical treatment is performed at a known service rate $\mu_1$. On treatment completion, patients are discharged back to the community, admitted to nursing homes, or referred to a geriatric ward (2–4) with routing probabilities $p_{1i}(t)$, $i = 2, 3, 4$, respectively. The number of beds in each geriatric ward $i$, $i = 2, 3, 4$, is $N_i$. If there are no available beds in the requested geriatric ward, its referred patients must wait in the hospital while blocking their current bed. This blocking mechanism is known as blocking after service (Balsamo et al. 2001). The treatment rates in Stations $i$, $i = 2, 3, 4$, are $\mu_i$. Frequently, the clinical condition of patients deteriorates while hospitalized in a geriatric ward, and they are hence readmitted to the hospital according to rate $\beta_i$, $i = 2, 3, 4$.

As mentioned, patients do die during their stay in a station, which we assume occurs at individual mortality rates $\theta_i$, $i = 1, 2, 3, 4$, for Stations 1–4. These mortality rates are significant and cannot be ignored. We follow the modeling of mortality as in Cohen et al. (2014), and in queueing theory parlance, we refer to it as "abandonments" that can occur while waiting or being treated. Although we use the same mortality rates while waiting and while being treated, if data prevail, our model can easily accommodate two different mortality rates per station.

### 3.2. Model Equations

We now introduce the functions $q_i(t)$, $i = 1, 2, 3, 4$, which denote the number of patients at Station $i$ at time $t$. The standard fluid modeling approach defines differential equations (DEs) describing the rate of change for each $q_i$. This direct approach leads to analytically intractable models that cannot not be justified as fluid limits of their corresponding stochastic counterparts. Moreover, these direct descriptions based on $q_i$ include indicator functions that are harder to analyze because of their discontinuity. Hence, we propose a new modeling approach, in which we introduce alternative functions $x_i(t)$, $i = 1, ..4$, that suffice to capture the state of the system. Then, we develop differential equations for $x_i$, which are tractable, and ultimately, we deduce $q_i$ from $x_i$.

This novel modeling approach also simplifies the convergence proof of the corresponding stochastic model, which is provided in Online Appendix B.

The value $x_1(t)$ denotes the number of arrivals to Station 1 who have not completed their treatment at Station 1 at time $t$. The values $x_i(t)$, $i = 2, 3, 4$, denote the number of patients who have completed treatment at Station 1 and require treatment at Station $i$ but have not yet completed their treatment at Station $i$ at time $t$ (these patients may still be blocked in Station 1). The dynamics of the system is captured through a set of DEs; each characterizes the rate of change in the number of patients at each state at time $t$. Let $\lambda_{total}(t)$ denote the arrival rate to Station 1 at time $t$ and $\delta_{total}(t)$ denote its departure rate. The DE for $x_1$ is, therefore,

$$\dot{x}_1(t) \triangleq \frac{dx_1(t)}{dt} = \lambda_{total}(t) - \delta_{total}(t). \tag{1}$$

Patients arrive to Station 1 from two sources: externally, according to rate $\lambda(t)$, and internally from Stations 2–4. Because $\beta_i$ is the readmission rate from Station $i$ back to Station 1, the internal arrival rate to Station 1 is $\sum_{i=2}^{4} \beta_i(x_i(t) \wedge N_i)$, where $x \wedge y = \min(x, y)$; here, $(x_i(t) \wedge N_i)$ denotes the number of patients in treatment at Station $i$. The total arrival rate to Station 1 at time $t$ is, therefore,

$$\lambda_{total}(t) = \lambda(t) + \sum_{i=2}^{4} \beta_i(x_i(t) \wedge N_i). \tag{2}$$

The total departure rate, $\delta_{total}(t)$, consists of two types. The first is owing to patients who die at an individual mortality rate $\theta_1$. Because patients might die while being hospitalized or waiting in queue, the rate at which patients die is $\theta_1 x_1(t)$. Let $b_i(t) = (x_i(t) - N_i)^+$, $i = 2, 3, 4$, denote the number of blocked patients in Station 1 at time $t$ (waiting for an available bed in Station $i$). If data are distinguished between different mortality rates while waiting ($\theta_{1q}$) versus being treated ($\theta_{1t}$), then the total mortality rate from Station 1 would be

$$\theta_{1q}\left[x_1(t) - \left(N_1 - \sum_{i=2}^{4} b_i(t)\right)\right]^+ + \theta_{1t}\left[x_1(t) \wedge \left(N_1 - \sum_{i=2}^{4} b_i(t)\right)\right];$$

here, the first addend represents the mortality rate while waiting for Station 1, and the second addend represents the mortality rate while being treated in Station 1. Note that the number of unblocked beds at Station 1 is $\left(N_1 - \sum_{i=2}^{4} b_i(t)\right)$, which can vary from zero to $N_1$.

The second departure type, $\delta_r(t)$, is of patients who complete their treatment at Station 1. The rate at which patients complete their treatment in Station 1 is

$$\delta_r(t) = \mu_1\left[x_1(t) \wedge \left(N_1 - \sum_{i=2}^{4} b_i(t)\right)\right], \tag{3}$$

where the expression in the brackets indicates the number of occupied unblocked beds at Station 1. Thus, the total departure rate at time $t$ is

$$\delta_{total}(t) = \theta_1 x_1(t) + \delta_r(t). \tag{4}$$

Using similar principles, we construct the DEs for the rate of change in $x_i$, $i = 2, 3, 4$. The referral rate to Station $i$ is $p_{1i}(t)$ multiplied by $\delta_r(t)$, the rate at which patients complete their treatment at Station 1. The departure rate of patients who have completed service at Station 1 but not at Station $i$ at time $t$ consists of the mortality rate $\theta_i x_i(t)$, readmission rate back to the hospital $\beta_i(x_i(t) \wedge N_i)$, and treatment completion rate $\mu_i(x_i(t) \wedge N_i)$.

The set of DEs for $x_i$, $i = 1, 2, 3, 4$, is, therefore,

$$\begin{aligned} \dot{x}_1(t) &= \lambda_{total}(t) - \delta_{total}(t), \\ \dot{x}_i(t) &= p_{1i}(t) \cdot \delta_r(t) - (\mu_i + \beta_i)(x_i(t) \wedge N_i) \\ &\quad - \theta_i x_i(t), \quad i = 2, 3, 4. \end{aligned} \tag{5}$$

The functions $q_i(t)$, $i = 1, 2, 3, 4$, which denote the number of patients at Station $i$ at time $t$, are

$$\begin{aligned} q_1(t) &= x_1(t) + \sum_{i=2}^{4} b_i(t); \\ q_i(t) &= x_i(t) \wedge N_i, \quad i = 2, 3, 4. \end{aligned} \tag{6}$$

The validation of the model, against both data and a discrete event stochastic simulation with different treatment distributions, is detailed in Online Appendix A. It shows that there is an excellent fit between the fluid model, the actual data, and the corresponding simulation results.

## 4. The Bed Allocation Model

The decision maker in our analysis is an organization that operates both hospitals and geriatric institutions. The objective is to find the optimal number of beds for each geriatric ward to minimize overall long-term underage and overage costs of care (beds) in the system.

Minimizing overage and underage costs is a typical objective in resource allocation problems (Porteus 2002). In our context, overage costs are incurred when geriatric beds remain empty while medical equipment, supply, and labor costs are still being paid. Labor costs in rehabilitation, for example, include, in addition to the cost of doctors and nurses, other professionals as well, such as neurophysiotherapists, orthopedic physical therapists, and occupational therapists. We denote by $C_o$ the per bed per day overage cost: this is the amount that could have been saved if the level of geriatric beds had been reduced by one unit in the event of an overage. This cost includes the per day per bed cost required for operating a geriatric bed. Underage cost, $C_u$, is incurred when patients are delayed in the hospital because of a lack of availability in the geriatric wards. Thus, it is the amount that could have been saved if the level of geriatric beds had been increased by one unit in the event

of an underage; $C_u$ is hence the per bed per day cost of hospitalization in hospitals minus the per bed per day cost in geriatric institutions. To elaborate, hospitalization costs also include risk costs, which are incurred when a patient is required to remain hospitalized. These costs include expected costs of patient medical deterioration of not providing the proper medical treatment (e.g., rehabilitation) and by exposing the patient to diseases and contaminations prevalent in hospitals. The sum of $C_o$ and $C_u$, which will later appear in the optimal solution in (15), amounts to the per bed per day hospitalization cost in hospitals. Excluding or underestimating the cost of risk will yield a lower bound for the required number of beds. Because our solution serves as a guide for thinking, meaningful insights can already be derived from such a lower bound.

We denote by $C_{o_i}$ and $C_{u_i}$ the overage and underage costs, respectively, for Stations $i$, $i = 2, 3, 4$. The resulting overall cost for Stations 2–4 over a planning horizon $T$ is

$$C^{(0)}(N_2, N_3, N_4) = \sum_{i=2}^{4} C^{(0)}(N_i), \tag{7}$$

where $C^{(0)}(N_i)$ is the total overage and underage costs for each Station $i$ given by

$$C^{(0)}(N_i) = \int_0^T \left[C_{u_i} \cdot b_i(t) + C_{o_i} \cdot (N_i - q_i(t))^+\right] \mathrm{d}t, \quad i = 2, 3, 4. \tag{8}$$

The first integrand is the underage cost calculated by adding up the number of blocked patients, and the second integrand is the overage cost calculated via the total number of vacant beds. Minimizing (7) will yield a constant capacity level for each geriatric ward over the whole planning horizon. In Section 6.2, we introduce a periodic reallocation of beds, which yields several capacity levels for each ward during the planning horizon.

**Remark 1.** Calculating the cost from (7) and (8) requires forecasting the arrival rate $\lambda(t)$ for the planning horizon $[0, T]$. This is done by using historical data: they show that there is an annual arrival rate pattern that repeats itself, whereas the volume increases at a rather constant rate each year. Hence, our healthcare partners can accurately predict the arrival rate over the planning horizon.

Minimizing (7), subject to (2)–(6), is analytically intractable, because $q_i(t)$ and $b_i(t)$ are solutions of a complex system of differential equations. To estimate the total cost, we use an offered load approximation to the time-varying demand for beds (Jennings et al. 1997, Whitt 2007). Thus, in Section 4.2, we present a closed form solution for minimizing the total underage and overage costs based on the offered load. Then, in Section 5.2, we compare our closed form solution with a numerical solution of the original problem.

### 4.1. Offered Loads in Our System

Given a resource, its *offered load* $r = \{r(t), t \geq 0\}$ represents the average amount of work being processed by that resource at time $t$ under the assumption that waiting and processing capacities are ample (no one queues up before service). In our context, offered load analysis is important for understanding demand. Indeed, we express demand in terms of patient bed days per day for the geriatric wards to determine appropriate bed capacity levels.

The calculation of the offered load is carried out by solving (5) (and (2)–(4)) with an unlimited capacity in Stations 2–4 ($N_i \equiv \infty$, $i = 2, 3, 4$). (Note that $b_i(t) \equiv 0$ for $i = 2, 3, 4$, which means that no patients are blocked.) These conditions yield the following set of DEs for the offered load $r_i$, $i = 1, \ldots, 4$ (just substitute $r_i$ for $x_i$ in (5)):

$$\begin{aligned} \dot{r}_1(t) &= \lambda(t) + \sum_{i=2}^{4} \beta_i r_i(t) - \theta_1 r_1(t) - \mu_1(r_1(t) \wedge N_1), \\ \dot{r}_i(t) &= p_{1i}(t) \cdot \mu_1(r_1(t) \wedge N_1) - (\beta_i + \theta_i + \mu_i) r_i(t), \\ & \quad i = 2, 3, 4. \end{aligned} \tag{9}$$

### 4.2. Estimating the Optimal Number of Beds Based on the Offered Load

The estimated overall cost for Stations 2–4 based on the offered load over the planning horizon $T$ is

$$C(N_2, N_3, N_4) = \sum_{i=2}^{4} C(N_i); \tag{10}$$

here, $C(N_i)$ is the underage plus overage cost for Station $i$ given by

$$C(N_i) = \int_0^T \left[C_{u_i} \cdot (r_i(t) - N_i)^+ + C_{o_i} \cdot (N_i - r_i(t))^+\right] \mathrm{d}t, \quad i = 2, 3, 4. \tag{11}$$

The first integrand corresponds to the underage cost, which is calculated by multiplying $C_{u_i}$ with the (proxy for) bed shortage $(r_i(t) - N_i)^+$ and integrating it over the planning horizon. The second integrand, the overage cost, is obtained by multiplying $C_{o_i}$ with the proxy for bed surplus $(N_i - r_i(t))^+$ and integrating it over the planning horizon as well.

**Remark 2.** How are these two proxies motivated?

First, under bed shortage (at cost $C_{u_i}$ per bed), we substitute $r_i$ for $x_i$. Second, under bed surplus (at cost $C_{o_i}$ per bed), we substitute $r_i$ for $q_i$. Third, because practically, $C_{u_i} \gg C_{o_i}$ (Section 5.1), the optimal solution must amplify reducing the number of blocked patients; hence, the more significant cost is incurred by bed surplus. Fourth, for calculating the latter cost and according to the offered load definition, $q_i \approx r_i$ when the system is underloaded. Under optimal bed allocation, bed blocking occurs infrequently enough so

that the load on each station is close to its offered load. Indeed, comparing the solutions according to the fluid model, its offered load approximation, and simulation results (Section 5.2) shows an excellent fit. Additional demonstration for the high quality of the offered load approximation is given in Online Appendix C.

The offered load for each station is a known function of $t$ that depends solely on input parameters but not on $N_2, N_3, N_4$. Thus, minimizing (10) is, in fact, a separable problem, which can be solved for each station separately. (When doing so below, we shall omit the $i$ in (11) for simplicity of notation.)

To minimize $C(N)$, we adopt the approach of Jennings et al. (1997) and treat $N$ as a continuous variable. We let $r_d = \{r_d(t) \mid 0 \le t \le T\}$ denote the *decreasing rearrangement* of $r$ on the interval $[0, T]$: $r_d$ on $[0, T]$ is characterized by being the unique decreasing function such that, for all $x \ge 0$, we have

$$\int_0^T 1_{\{r(t)\ge x\}}\mathrm{d}t = \int_0^T 1_{\{r_d(t)\ge x\}}\mathrm{d}t; \tag{12}$$

here, $1_{\{r(t)\ge x\}}$ denotes the indicator function for the event $\{r(t) \ge x\}$. Existence and uniqueness of $r_d$ were established in Hardy et al. (1952). The interpretation of Equation (12) is that both $r(t)$ and $r_d(t)$ spend the same amount of time above and under any level $x$. We can now rewrite $C(N)$ as follows:

$$\begin{aligned}
C(N) &= \int_0^T [C_u \cdot (r(t) - N)^+ + C_o \cdot (N - r(t))^+]\mathrm{d}t \qquad (13)\\
&= \int_N^\infty C_u \int_0^T 1_{\{r(t)\ge x\}}\mathrm{d}t\,\mathrm{d}x + \int_0^N C_o \int_0^T 1_{\{r(t)\le x\}}\mathrm{d}t\,\mathrm{d}x\\
&= \int_0^\infty C_u \int_0^T 1_{\{r(t)\ge x\}}\mathrm{d}t\,\mathrm{d}x - \int_0^N C_u \int_0^T 1_{\{r(t)\ge x\}}\mathrm{d}t\,\mathrm{d}x\\
&\quad + \int_0^N C_o\left[T - \int_0^T 1_{\{r(t)\ge x\}}\mathrm{d}t\right]\mathrm{d}x\\
&= \int_0^\infty C_u \int_0^T 1_{\{r(t)\ge x\}}\mathrm{d}t\,\mathrm{d}x - \int_0^N (C_u + C_o)\\
&\quad \cdot \int_0^T 1_{\{r(t)\ge x\}}\mathrm{d}t\,\mathrm{d}x + C_o TN\\
&= \int_0^\infty C_u \int_0^T 1_{\{r_d(t)\ge x\}}\mathrm{d}t\,\mathrm{d}x - \int_0^N (C_u + C_o)\\
&\quad \cdot \int_0^T 1_{\{r_d(t)\ge x\}}\mathrm{d}t\,\mathrm{d}x + C_o TN,
\end{aligned}$$

where the first equality is achieved by substituting

$$\begin{aligned}
(r(t) - N)^+ &= \int_N^\infty 1_{\{r(t)\ge x\}}\mathrm{d}x,\\
(N - r(t))^+ &= \int_0^N 1_{\{r(t)\le x\}}\mathrm{d}x
\end{aligned} \tag{14}$$

and interchanging the order of integration.

We are now ready for Theorem 1, which identifies the optimal number of beds, $N^*$. The proof of the theorem is provided in Online Appendix D. Note that our proof does not require that $r(t)$ and $\lambda(t)$ be continuous or differentiable. (These assumptions were needed in Jennings et al. (1997).)

**Theorem 1.** *The number of beds that minimizes $C(N)$ is given by*

$$N^* = r_d\left(\frac{C_o T}{C_o + C_u}\right). \tag{15}$$

In Online Appendix E, we explain how $N^*$ arose as a candidate for minimizing $C(N)$.

**Remark 3.** Alternatively, one can obtain the solution by building the cumulative relative frequency function for $r$ and noting the similarity between our problem and the Newsvendor problem (Arrow et al. 1951, Nahmias and Cheng 2009) for inventory management. In this case, we interpret the frequency as probability. This approach is similar to the reduction to the Newsvendor problem in Harrison and Zeevi (2005). However, our solution in (15) is more natural (more directly related to the time-varying nature of our models and their underlying systems); more importantly, this time-varying view naturally enables the solution of two extensions: setup cost per new bed (Section 6.1) and periodic reallocation of beds (Section 6.2) (these are beyond the scope of the Newsvendor problem extension). Note that, in the case of constant arrival rates, the offered load would also be constant, and the optimal number of beds would exactly equal the offered load.

## 5. Numerical Results

In this section, we apply our model to data to validate our solution (Sections 5.1 and 5.2), calculate the imputed costs (Section 5.3), and provide structural insights and managerial recommendations (Section 5.4).

### 5.1. An Illustrative Example

Our healthcare partners were willing to share with us some of their financial reports and cost data. Rigorous calculations based on these data (some of which are confidential) yielded the following critical fractiles required for (15). The hospitalization cost in mechanical ventilation wards is the highest among the geriatric wards, and as it turns out, $C_{u_3} = 1.882C_{o_3}$. In rehabilitation wards, the ratio is $C_{u_2} = 2.667C_{o_2}$, because the hospitalization there is less expensive. Finally, the ratio for skilled nursing care is $C_{u_4} = 4.267C_{o_4}$, because the hospitalization cost there is the lowest among the geriatric wards.

We used the fluid model developed in Section 3 together with our two-year historical data to forecast the offered load for a subsequent three-year planning horizon, where the demand for beds (e.g., the arrival rate) increases every year because of population aging and growth. Then, by using Matlab, we numerically

constructed the functions $r_d$ for each ward (by sorting the function values of $r$). The optimal number of beds is the value of these functions at the critical point as in (15). Because the value of $N^*$ is not necessarily an integer, it must be rounded. Rounding up versus down has minor significance, because the solution here serves as a guide for a large organization that provides healthcare services for an entire district. Therefore, our solution provides insights regarding the difference between the suggested allocation and the current capacity.

Figure 3, left panel presents the optimal number of beds (the dashed lines) compared with the offered load (solid lines). The optimal number of beds for each ward was calculated by rounding up the result from Equation (15). The optimal solution implies increasing the current number of beds by 25%, 35%, and 33% in rehabilitation, mechanical ventilation, and skilled nursing care, respectively. In total, this is an increase to 577 beds from the current 439 beds. This will lead to overage and underage cost reductions of 51%, 53%, and 69%, respectively; here, we compared the cost using our solution with the current number of beds for the same arrival forecast. We believe that there are two major reasons for this dramatic cost reduction. The first is the lack of a model in practice, such as the one introduced here: such a model would take blocking and its related costs into account, which would guide planners. The second reason is the difficulties in increasing the present budget toward acquiring new beds. We provide more details and calculate imputed costs in Section 5.3.

Figure 3, right panel presents the waiting list length of each geriatric ward under the optimal number of beds. Note that the waiting lists are shorter (compared with the current situation presented in Figure 2) by 67%, 74%, and 88% in rehabilitation, mechanical ventilation, and skilled nursing care, respectively. This occurs, although shortening the waiting lists is not directly included in our objective function. Indeed, we aimed at minimizing overage and underage costs; because blocking costs are significant, reducing the total cost is achieved by reducing blocking, which in turn, leads to significant shorter waiting lists.

### 5.2. Solution Validation and Cost Comparison

In addition to validating our fluid model against data and stochastic simulation results (Online Appendix A), in this section, we validate our bed planning solution.

Thus far, two cost functions were presented for estimating the optimal number of geriatric beds. The first, $C^{(0)}(N_2, N_3, N_4)$ in (7), is based on the time-varying number of patients as derived from the solution of the fluid equations in (6). Because minimizing $C^{(0)}(N_2, N_3, N_4)$ is analytically intractable, we introduce the second cost function, $C(N_2, N_3, N_4)$ in (10), which estimates the total cost based on an offered load approximation to the time-varying demand for beds.

To validate the approximated cost function, we compared the optimal solutions for the two problems with the optimal solution derived from our stochastic simulation model. In the latter, the arrivals, duration times, and routing percentages are random variables (Online Appendix A). All parameters, including the size of the system, are realistic for the system that we analyze.

The solution for $C(N_2, N_3, N_4)$ was calculated by our closed form expression in (15). The solution for $C^{(0)}(N_2, N_3, N_4)$ was achieved by numerically solving the optimization problem in (7) and (8); this was done by solving the fluid model in (5) and (6) for each capacity combination, calculating the total cost according to (7), and choosing the capacity combination with the minimal cost. Finally, the solution for the stochastic simulation model was achieved by calculating, for each capacity combination, the total underage and overage costs. This was done by using (7) and (8), where instead of $q_i$ and $b_i$, $i = 2, 3, 4$, we used the corresponding numbers from the simulation results. Then, we chose the combination that minimized the cost. In other words, the solutions according to $C^{(0)}(N_2, N_3, N_4)$ and simulation were carried out by a three-dimensional search (over $N_2$, $N_3$, and $N_4$). Table 1 summarizes this comparison by presenting the optimal number of beds and the optimal cost according to each method. In addition, we calculated the differences in percentages between the two methods for each ward separately and then, all together. The last column in Table 1 presents the maximal difference between the solutions. The maximal difference varied from 1% to 1.6% when comparing bed allocations and from 1.1% to 3.4% when comparing total cost. This excellent fit is typical; indeed, we obtained similar differences when comparing the three solutions under several other scenarios of overage and underage costs.

### 5.3. The Imputed Overage and Underage Costs

In addition to estimating the $C_o/C_u$ ratio given to us by our healthcare organization, it is of interest to examine $C_o$ and $C_u$ as imputed costs. These imputed costs are based on observed decisions that, in our case, are the numbers of beds that decision makers allocate to each geriatric ward. To this end, we use the current number of beds in each geriatric ward to extract the model's parameters $C_o$ and $C_u$, or more accurately, the ratio $C_o/C_u$. (A similar approach was taken by Olivares et al. 2008.) Suppose that the current allocation $N$ is optimal; we then define

$$r_d^{-1}(N) \equiv \sup\{t \mid r_d(t) \geq N\} \tag{16}$$

**Table 1.** Comparing Optimal Solutions (Number of Beds and Overage and Underage Costs per Year)—$C^{(0)}(N_2, N_3, N_4)$ Vs. $C(N_2, N_3, N_4)$ Vs. Simulation

| | $N^*$ (total cost) | | | |
|---|---|---|---|---|
| Ward | $C^{(0)}(N_2, N_3, N_4)$ | $C(N_2, N_3, N_4)$ | Simulation | Maximal difference, % |
| Rehabilitation | 295 (2,601,667) | 292 (2,683,042) | 294 (2,633,167) | 1.0 (3.0) |
| Mechanical ventilation | 128 (1,493,917) | 126 (1,547,000) | 128 (1,499,167) | 1.6 (3.4) |
| Skilled nursing | 161 (1,213,333) | 159 (1,226,750) | 160 (1,215,667) | 1.3 (1.1) |
| Total number of beds | 584 (5,308,917) | 577 (5,456,792) | 582 (5,348,000) | 1.2 (2.7) |

as the time during which underage costs were incurred. Let $I$ denote the fraction of time during which underage costs were incurred. Consequently, from Theorem 1, we have

$$I = \frac{r_d^{-1}(N)}{T} = \frac{C_o}{C_o + C_u}. \tag{17}$$

We now present our data as a sequence of $n$ days: $(t_i, r(t_i))$ for $i = 1, \ldots, n$, where $t_i$ denotes a single time point for day $i$. Then, we define $\bar{I}$ to be an estimator for the fraction of time during which underage costs were incurred:

$$\bar{I} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{r(t_i) \geq N\}}. \tag{18}$$

We replace $r_d^{-1}(N)/T$ with $\bar{I}$ in (17) to get

$$\bar{I} = \frac{C_o}{C_o + C_u}. \tag{19}$$

According to our data, $\bar{I}_2 = 0.74$ in rehabilitation, $\bar{I}_3 = 0.91$ in skilled nursing care, and $\bar{I}_4 = 1$ in mechanical ventilation. Therefore, the imputed costs are $C_{u_2} = 0.35C_{o_2}$ (versus $C_{u_2} = 2.667C_{o_2}$ according to the financial reports) in rehabilitation, $C_{u_3} = 0.099C_{o_3}$ (versus $C_{u_3} = 1.882C_{o_3}$) in skilled nursing care, and $C_{u_4} = 0$ (versus $C_{u_4} = 4.267$) in mechanical ventilation. The differences in the imputed costs among the three wards are caused by different hospitalization costs as explained in Section 5.1.

There is a big difference between the ratio $C_u/C_o$ according to the financial reports and according to the imputed costs. This may imply that blocking costs are neglected or underestimated when determining the geriatric bed capacity. Another possible explanation is that, although there is a central decision maker that owns both the hospitals and geriatric institutions, decisions are locally optimized.

Note that the financial estimations of overage and underage costs are adequate for situations close to the current one. Adding hundreds of beds will require additional investment (e.g., real estate) that is not captured by the current estimates. The case where the marginal cost per bed is higher because of bed setup is addressed in Section 6.1.

### 5.4. Managerial Insights for the Optimal Solution

The function $r_d$ in the optimal solution (15) is decreasing in $[0, T]$. As explained earlier, the ratio $C_o/(C_o + C_u)$ in the optimal solution is the hospitalization cost ratio between a geriatric bed and a hospital bed. As the gap between these two costs widens, more geriatric beds will be needed. Indeed, in Figure 3, the optimal number of beds in skilled nursing care is relatively high compared with the offered load. The reason for this is the relatively low hospitalization cost in this ward. In mechanical ventilation, however, the optimal number of beds is relatively low compared with the offered load, because the hospitalization cost there is higher.

Note that, when the optimal allocation provided in (15) is too large to be implemented at once, it can also be implemented gradually while estimating the cost reduction for each step according to (10). In addition, given a specific budgetary constraint, our approach allows one to evaluate the cost of (or numerically seek an optimal) bed capacity.

Figure 3 shows long periods of low bed occupancy, especially in skilled nursing care and rehabilitation. To accommodate for the seasonal demand, we seek a more flexible solution, such as reallocating beds between wards. To this end, we first calculate the total offered load for the three wards; then, we minimize (11) to find the total required number of beds. The optimal solution will then require fewer beds overall (566 beds instead of 577), but it will lead to only an additional decrease of 5% in the total cost. The improvement is relatively modest because of the correlated patterns of the offered load among the wards; this implies that more beds are needed in all three wards at the same time. Thus, reallocating beds between wards is less effective in reducing the cost.

Consequently, a more flexible and responsive policy to fluctuations in demand can be achieved by adding and removing beds throughout the year. Because the costs that we consider are for staffed beds (namely beds for which there is an assigned medical staff) regardless of their occupancy, not staffing beds during overage periods would reduce the costs. According to our healthcare partners, implementing two bed capacity levels per year, which by our model, implies two capacity switches each year, is feasible. For example, it is possible to open a specific area/ward when demand is

high (usually in the winter) and close this area when demand is low (usually in the summer). The described policy is feasible, because most "bed cost" is related to labor cost and medical supplies; the latter can be purchased seasonally, whereas the former can be changed because of the existing flexibility of staffing levels (e.g., reallocating workers within facilities in the same organization or changing the work load of part-time workers throughout the year). We formally introduce and analyze the periodic reallocation problem in Section 6.2.

## 6. Extensions

In this section, we present two extensions to our model. The first extension, at the strategic level, adds setup costs for allocating new beds. The second extension, at the operational level, allows a periodic reallocation of beds.

### 6.1. Including Setup Cost per New Bed

In this section, we analyze a case where there is a fixed setup cost, $K$, associated with the introduction of each new bed. The setup cost may be associated with recruitment and training of new staff or the purchase of new equipment. We assume that the setup cost may vary with bed types. Let $B$ denote the current bed capacity; then, the overall cost for a geriatric ward is

$$C_K(N) = C(N) + K(N - B)^+, \tag{20}$$

where $C(N)$ is the overall cost analyzed in Section 4 and $(N - B)^+$ is the number of new beds. The planning horizon, $T$, reflects an organizational policy regarding investments and hence, should be long enough for an investment in new beds to be worthwhile.

**Theorem 2.** *The optimal number of beds that minimizes $C_K(N)$ is given by*

$$N_k^* = \begin{cases} r_d\left(\dfrac{C_o T}{C_o + C_u}\right), & \text{if} \quad r_d\left(\dfrac{C_o T}{C_o + C_u}\right) \le B \\ r_d\left(\dfrac{C_o T + K}{C_o + C_u}\right), & \text{if} \quad r_d\left(\dfrac{C_o T + K}{C_o + C_u}\right) \ge B \\ B, & \text{otherwise.} \end{cases} \tag{21}$$

We prove Theorem 2 in Online Appendix F.

Note that $r_d(\cdot)$ is defined on the interval $[0, T]$; hence, when $C_u T < K$, then $r_d(\cdot)$ is undefined, because

$$\frac{C_o T + K}{C_o + C_u} > \frac{C_o T + C_u T}{C_o + C_u} = T.$$

In this case, only the first condition of $N_K^*$ is relevant. Therefore, the solution will not include the introduction of new beds. An intuitive explanation is that, for a high bed setup cost, it may be preferable to pay the underage cost for the entire planning horizon.

When implementing the method described in Section 5.3, we get that $K$, the imputed setup cost per new bed, follows the condition $K \ge T(\bar{I}C_u - (1 - \bar{I})C_o)$; here, $\bar{I}$ is the estimator for the fraction of time during which underage costs were incurred (Section 5.3). Under the current financial estimations and a five-year planning horizon, we get that the imputed setup costs are $K_2 \ge 620,600$; $K_3 \ge 587,940$; and $K_4 \ge 1,596,970$. These costs are about three times higher than the setup costs according to our healthcare partners. This implies that, even when considering setup costs, the blocking costs are underestimated when determining geriatric bed capacities.

### 6.2. Periodic Reallocation of Beds

Managers of geriatric institutions acknowledge that it is feasible to change the number of beds during the year to compensate for seasonal variations in demand. Note that changing the number of beds also implies changing staff levels (which are typically proportional to the number of beds) and other related costs. The planning horizon remains the same, but we divide each year into several periods. We assume that labor can be flexible but only to a certain degree (e.g., staffing levels can be adjusted twice a year but not on a daily/weekly basis). We then determine the preferable periods (location and length) and the number of beds required for each period. For example, an optimal reallocation policy would determine a certain capacity during the first three months and the last two months of every year in the planning horizon and possibly, a different capacity during the seven other months of every year. To this end, we introduce a reallocation cost, $C_r$, associated with adding and removing a bed.

Because of feasibility constraints from our partner hospital chain, we allow only two capacity levels throughout the planning horizon. Nevertheless, the methodology that we present can be implemented in other settings where more capacity levels are possible. Moreover, because of the nature/shape of the demand, having two capacity levels corresponds to changing capacity levels twice each year.

Let $\mathcal{T} = [0, T]$ denote the planning horizon interval, and let $\mathcal{I}$ denote the time interval (location and length) in which there are $N_{\mathcal{I}}$ geriatric beds (in $\mathcal{T}\backslash\mathcal{I}$, there are $N_{\mathcal{T}\backslash\mathcal{I}}$ geriatric beds). Our objective is to find $\mathcal{I}$, $N_{\mathcal{I}}$, and $N_{\mathcal{T}\backslash\mathcal{I}}$ that minimize the total underage and overage costs.

To this end, we split $r(t)$ into two functions: $r_{\mathcal{I}}(t)$ for the capacity level in $\mathcal{I}$ and $r_{\mathcal{T}\backslash\mathcal{I}}(t)$ for the capacity level in $\mathcal{T}\backslash\mathcal{I}$. The functions $r_{\mathcal{I}}(t)$ and $r_{\mathcal{T}\backslash\mathcal{I}}(t)$ are defined on the intervals $[0, |\mathcal{I}|]$ and $[0, |\mathcal{T}\backslash\mathcal{I}|]$, respectively, by concatenating the relevant intervals from $r(t)$ and shifting the functions to $t = 0$. We define the functions $r_{d_{\mathcal{I}}}(t)$ and $r_{d_{\mathcal{T}\backslash\mathcal{I}}}(t)$ to be the decreasing rearrangements of $r_{\mathcal{I}}(t)$ and $r_{\mathcal{T}\backslash\mathcal{I}}(t)$, respectively, exactly as we defined

$r_d(t)$ in Section 4. The total underage and overage costs are, therefore,

$$\begin{aligned} C(\mathcal{I}, N_{\mathcal{I}}, N_{\mathcal{T}\setminus\mathcal{I}}) &= C(\mathcal{I}, N_{\mathcal{I}}) + C(\mathcal{T}\setminus\mathcal{I}, N_{\mathcal{T}\setminus\mathcal{I}}) \\ &\quad + C_r\left|N_{\mathcal{T}\setminus\mathcal{I}} - N_{\mathcal{I}}\right| \\ &= \int_{\mathcal{I}}\left[C_u(r(t) - N_{\mathcal{I}})^+ + C_o(N_{\mathcal{I}} - r(t))^+\right]dt \\ &\quad + \int_{\mathcal{T}\setminus\mathcal{I}}\left[C_u(r(t) - N_{\mathcal{T}\setminus\mathcal{I}})^+ \right. \\ &\quad \left. + C_o(N_{\mathcal{T}\setminus\mathcal{I}} - r(t))^+\right]dt + C_r\left|N_{\mathcal{T}\setminus\mathcal{I}} - N_{\mathcal{I}}\right|, \end{aligned} \tag{22}$$

where $C(\mathcal{I}, N_{\mathcal{I}})$ and $C(\mathcal{T}\setminus\mathcal{I}, N_{\mathcal{T}\setminus\mathcal{I}})$ denote the overage and underage costs for intervals $\mathcal{I}$ and $\mathcal{T}\setminus\mathcal{I}$, respectively.

**Theorem 3.** *The number of beds that minimizes (22) for a fixed $\mathcal{I}$ is*

$$\begin{cases} N^*_{\mathcal{I}} = N^{\mathcal{I}}_-, \quad N^*_{\mathcal{T}\setminus\mathcal{I}} = N^{\mathcal{T}\setminus\mathcal{I}}_+, & \text{if} \quad N^{\mathcal{I}}_- \le N^{\mathcal{T}\setminus\mathcal{I}}_+, \\ N^*_{\mathcal{I}} = N^{\mathcal{I}}_+, \quad N^*_{\mathcal{T}\setminus\mathcal{I}} = N^{\mathcal{T}\setminus\mathcal{I}}_-, & \text{if} \quad N^{\mathcal{I}}_+ \ge N^{\mathcal{T}\setminus\mathcal{I}}_-, \\ N^*_{\mathcal{I}} = N^*_{\mathcal{T}\setminus\mathcal{I}} = N^*, \quad \text{as in (15)}, & \text{otherwise}. \end{cases} \tag{23}$$

*Here, $N^{\mathcal{A}}_{\pm} = r_{d_{\mathcal{A}}}\left(\frac{C_o|\mathcal{A}| \pm C_r}{C_o + C_u}\right)$ for every interval $\mathcal{A}$ (where $\mathcal{A}$ is either $\mathcal{I}$ or $\mathcal{T}\setminus\mathcal{I}$).*

We prove Theorem 3 in Online Appendix G.

Note that the option in the third line in (23) suggests determining only one capacity level (e.g., it is preferable not to reallocate beds throughout the planning horizon). In particular, because $r_{d_{\mathcal{I}}}(\cdot)$ and $r_{d_{\mathcal{T}\setminus\mathcal{I}}}(\cdot)$ are defined on the intervals $[0, |\mathcal{I}|]$ and $[0, |\mathcal{T}\setminus\mathcal{I}|]$, respectively, when $C_u|\mathcal{I}| > C_r$ or $C_u|\mathcal{T}\setminus\mathcal{I}| > C_r$, it is preferable to pay the underage cost for the entire period than to pay the reallocation cost, $C_r$.

**6.2.1. A Numerical Example.** We now solve the periodic reallocation problem for a three-year planning horizon. Figure 4 depicts the solutions for three cases. The solid lines in Figure 4 represent the offered load for each ward, whereas the dashed lines in Figure 4 represent the optimal number of beds. The first case (Figure 4, upper left panel) is when no reallocation costs are introduced ($C_r = 0$). This solution yields 35%, 22%, and 31% underage and overage cost reductions in rehabilitation, mechanical ventilation, and skilled nursing care, respectively, compared with the constant allocation. The second case (Figure 4, upper right panel) is when reallocation costs are introduced; in this case, the gap between the two capacity levels narrows. In particular, the optimal allocation in mechanical ventilation is constant, because it is not worthy to invest the reallocation cost (e.g., $C_r > C_u|\mathcal{I}|$ or $C_r > C_u|\mathcal{T}\setminus\mathcal{I}|$). The third case (Figure 4, lower right panel) presents the optimal periodic reallocation when four reallocation points are allowed and no reallocation costs are introduced. Figure 4, lower left panel presents the waiting list lengths for each ward under the optimal reallocation policy when no reallocation costs are introduced; this is in comparison with the current situation presented in Figure 2 and the constant allocation presented in Figure 3, right panel.

### 6.3. Managerial Recommendations on Extensions

The major cost reduction compared with the current situation for the three wards is achieved by adopting the proposed policy of a constant number of beds. Periodic allocations allow for extra cost reductions compared with the policy with a constant number of beds. Thus, a reasonable policy would be to adopt the constant allocation as a clear first step; then, if feasible, a periodic reallocation ought to be contemplated. In some cases, when the reallocation cost is higher than the underage period cost, it is preferable to remain with the constant allocation (see the case for mechanical ventilation ward in Figure 4, upper right panel).

Another option, which can help reduce the load, is to divert more geriatric patients in peak periods to home healthcare services or virtual hospitals rather than to geriatric institutions (Ticona and Schulman 2016). In this case, multidisciplinary home teams treat the patient at home rather than in the hospital. Home care hospitalization was found to be as effective, be less expensive, be shorter duration, and increase patient satisfaction compared with the same treatment received in a hospital (Shepperd et al. 2008, Caplan et al. 2012). Moreover, according to our analysis, even a 10% diversion of patients requiring geriatric hospitalization to home care will reduce the overage and underage costs by about 25% on average and will shorten the waiting lists in hospital by 30% on average.

## 7. Future Research

There are multiple directions worthy of future research, and two of which will now be described. The first is to modify the structure of the system by adding an intermediate ward (i.e., a stepdown unit) for subacute geriatrics (Wolstenholme 1999) between the hospital and the geriatric institutions. Such an intermediate ward would be designated for elderly patients with an expected long stay in the hospital before continuing on to a geriatric ward. Adding a subacute ward can both reduce the workload and bed occupancy in hospitals and improve the patient flow in and out of the hospital.

The second direction is a capacity allocation problem in which, given a predefined budget, the planners must decide where it is most beneficial to add new beds: in hospitals, intermediate wards, or geriatric wards. The simple version of this question (without intermediate wards), in fact, triggered this research.

### Acknowledgments

The authors would like to thank the editorial team for their thoughtful comments and support that helped to improve the paper.

## References

Afèche P, Araghi M, Baron O (2017) Customer acquisition, retention, and queueing-related service quality: Optimal advertising, staffing, and priorities for a call center. *Manufacturing Service Oper. Management* 19(4):509–712.

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

Armony M, Israelit S, Mandelbaum A, Marmor YN, Tseytlin Y, Yom-Tov GB (2015) On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* 5(1):146–194.

Arrow KJ, Harris T, Marschak J (1951) Optimal inventory policy. *Econometrica J. Econometric Soc.* 19(3):250–272.

Ata B, Killaly BL, Olsen TL, Parker RP (2013) On hospice operations under medicare reimbursement policies. *Management Sci.* 59(5): 1027–1044.

Balsamo S, de Nitto Personé V, Onvural R (2001) *Analysis of Queueing Networks with Blocking* (Springer, New York).

Bassamboo A, Harrison JM, Zeevi A (2006) Design and control of a large call center: Asymptotic analysis of an lp-based method. *Oper. Res.* 54(3):419–435.

BBC News (2016) Hospital bed-blocking costs NHS England £900m a year. Accessed February 5, 2016, http://www.bbc.com/news/health-35481849.

Bekker R, de Bruin AM (2010) Time-dependent analysis for refused admissions in clinical wards. *Ann. Oper. Res.* 178(1):45–65.

Bretthauer KM, Heese HS, Pun H, Coe E (2011) Blocking in healthcare operations: A new heuristic and an application. *Production Oper. Management* 20(3):375–391.

Caplan GA, Sulaiman NS, Mangin DA, Ricauda NA, Wilson AD, Barclay L (2012) A meta-analysis of "hospital in the home." *Medical J. Australia* 197(9):512–519.

Cheah JY, Smith JM (1994) Generalized M/G/C/C state dependent queueing models and pedestrian traffic flows. *Queueing Systems* 15(1–4):365–386.

Cochran JK, Bharti A (2006) Stochastic bed balancing of an obstetrics hospital. *Health Care Management Sci.* 9(1):31–45.

Cohen I, Mandelbaum A, Zychlinski N (2014) Minimizing mortality in a mass casualty event: Fluid networks in support of modeling and staffing. *IIE Trans.* 46(7):728–741.

De Vries T, Beekman RE (1998) Applying simple dynamic modelling for decision support in planning regional health care. *Eur. J. Oper. Res.* 105(2):277–284.

Desai MS, Penn ML, Brailsford S, Chipulu M (2008) Modelling of hampshire adult services—gearing up for future demands. *Health Care Management Sci.* 11(2):167–176.

Donoghue WF (1969) *Distributions and Fourier Transforms* (Elsevier, Amsterdam).

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 Management Sci.* 1(2):143–149.

Faddy M, Graves N, Pettitt A (2009) Modeling length of stay in hospital and other right skewed data: Comparison of phase-type, gamma and log-normal distributions. *Value Health* 12(2):309–314.

Faddy MJ, McClean SI (2005) Markov chain modelling for geriatric patient care. *Methods Inform. Medicine* 44(3):369–373.

Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. *Manufacturing Service Oper. Management* 4(3):208–227.

Gershwin SB (1987) An efficient decomposition method for the approximate evaluation of tandem queues with finite storage space and blocking. *Oper. Res.* 35(2):291–305.

Gray LC, Broe GA, Duckett SJ, Gibson DM, Travers C, McDonnell G (2006) Developing a policy simulator at the acute-aged care interface. *Australian Health Rev.* 30(4):450–457.

Green LV (2004) Capacity planning and management in hospitals. *Operations Research and Health Care* (Springer, New York), 15–41.

Green LV, Kolesar PJ, Whitt W (2007) Coping with time-varying demand when setting staffing requirements for a service system. *Production Oper. Management* 16(1):13–39.

Hardy GH, Littlewood JE, Pólya G (1952) *Inequalities* (Cambridge University Press, Cambridge, UK).

Hare WL, Alimadad A, Dodd H, Ferguson R, Rutherford A (2009) A deterministic model of home and community care client counts in British Columbia. *Health Care Management Sci.* 12(1):80–98.

Harrison GW, Millard PH (1991) Balancing acute and long-term care: The mathematics of throughput in departments of geriatric medicine. *Methods Inform. Medicine* 30(3):221–228.

Harrison JM, Zeevi A (2005) A method for staffing large call centers based on stochastic fluid models. *Manufacturing Service Oper. Management* 7(1):20–36.

Hillier FS, Boling RW (1967) Finite queues in series with exponential or erlang service times—a numerical approach. *Oper. Res.* 15(2): 286–303.

Jennings OB, Massey WA, McCalla C (1997) Optimal profit for leased lines services. *Proc. 15th Internat. Teletraffic Congress, Washington, DC*, vol. 15, 803–814.

Kao EPC, Tung GG (1981) Bed allocation in a public health care delivery system. *Management Sci.* 27(5):507–520.

Katsaliaki K, Brailsford S, Browning D, Knight P (2005) Mapping care pathways for the elderly. *J. Health Organ. Management* 19(1): 57–72.

Kerbache L, MacGregor Smith J (1987) The generalized expansion method for open finite queueing networks. *Eur. J. Oper. Res.* 32(3):448–461.

Kerbache L, MacGregor Smith J (1988) Asymptotic behavior of the expansion method for open finite queueing networks. *Comput. Oper. Res.* 15(2):157–169.

Koizumi N, Kuno E, Smith TE (2005) Modeling patient flows using a queuing network with blocking. *Health Care Management Sci.* 8(1):49–60.

Li A, Whitt W, Zhao J (2015) Staffing to stabilize blocking in loss models with time-varying arrival rates. *Probab. Engrg. Inform. Sci.* 30(2):185–211.

Liu Y, Whitt W (2011a) Large-time asymptotics for the$G_t/M_t/s_t + GI_t$ many-server fluid queue with abandonment. *Queueing Systems* 67(2):145–182.

Liu Y, Whitt W (2011b) A network of time-varying many-server fluid queues with customer abandonment. *Oper. Res.* 59(4):835–846.

Liu Y, Whitt W (2012) The$G_t/GI/s_t + GI$many-server fluid queue. *Queueing Systems* 71(4):405–444.

Liu Y, Whitt W (2014) Many-server heavy-traffic limit for queues with time-varying parameters. *Ann. Appl. Probab.* 24(1):378–421.

Mandelbaum A, Massey WA, Reiman MI (1998) Strong approximations for Markovian service networks. *Queueing Systems* 30(1–2): 149–201.

Mandelbaum A, Massey WA, Reiman MI, Rider B (1999) Time varying multiserver queues with abandonment and retrials. *Proc. 16th Internat. Teletraffic Conf., Edinburgh, Scotland*, vol. 4, 4–7.

Marazzi A, Paccaud F, Ruffieux C, Beguin C (1998) Fitting the distributions of length of stay by parametric models. *Medical Care* 36(6):915–927.

McCalla C, Whitt W (2002) A time-dependent queueing-network model to describe the life-cycle dynamics of private-line telecommunication services. *Telecomm. Systems* 19(1):9–38.

McClean S, Millard P (1993) Patterns of length of stay after admission in geriatric medicine: An event history approach. *Statistician* 42(3):263–274.

McClean S, Millard P (2006) Where to treat the older patient? Can Markov models help us better understand the relationship between hospital and community care? *J. Oper. Res. Soc.* 58(2):255–261.

Nahmias S, Cheng Y (2009) *Production and Operations Analysis* (McGraw-Hill, New York).

Namdaran F, Burnet C, Munroe S (1992) Bed blocking in Edinburgh hospitals. *Health Bull.* 50(3):223–227.

NHS England (2015) Bed availability and occupancy data. Accessed August 20, 2015, https://www.england.nhs.uk/statistics/statistical-work-areas/bed-availability-and-occupancy/bed-data-overnight/.

OECD iLibrary (2013) Health at a glance. Accessed November 21, 2013, http://dx.doi.org/10.1787/health_glance-2013-en.

Olivares M, Terwiesch C, Cassorla L (2008) Structural estimation of the newsvendor model: An application to reserving operating room time. *Management Sci.* 54(1):41–55.

Osorio C, Bierlaire M (2009) An analytic finite capacity queueing network model capturing the propagation of congestion and blocking. *Eur. J. Oper. Res.* 196(3):996–1007.

Perros HG (1994) *Queueing Networks with Blocking* (Oxford University Press, Oxford, UK).

Porteus EL (2002) *Foundations of Stochastic Inventory Theory* (Stanford University Press, Stanford, CA).

Rohleder TR, Cooke D, Rogers P, Egginton J (2013) Coordinating health services: An operations management perspective. *Handbook of Healthcare Operations Management* (Springer, Bring), 421–445.

Rubin SG, Davies GH (1975) Bed blocking by elderly patients in general-hospital wards. *Age Ageing* 4(3):142–147.

Shepperd S, Doll H, Angus RM, Clarke MJ, Iliffe S, Kalra L, Ricauda NA, Wilson AD (2008) Admission avoidance hospital at home. *Cochrane Database Syst Rev.* 4:CD007491.

Shi P, Chou MC, Dai JG, Ding D, Sim J (2015) Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Sci.* 62(1):1–28.

Takahashi Y, Miyahara H, Hasegawa T (1980) An approximation method for open restricted queueing networks. *Oper. Res.* 28(3 Part I):594–602.

Taylor G, McClean S, Millard P (1997) Continuous-time Markov models for geriatric patient behaviour. *Appl. Stochastic Models Data Anal.* 13(3–4):315–323.

Taylor GJ, McClean SI, Millard PH (2000) Stochastic models of geriatric patient bed occupancy behaviour. *J. Royal Statist. Soc. Ser. A* 163(1): 39–48.

Ticona L, Schulman KA (2016) Extreme home makeover—the role of intensive home health care. *New England J. Medicine* 375(18): 1707–1709.

Travers CM, McDonnell GD, Broe GA, Anderson P, Karmel R, Duckett SJ, Gray LC (2008) The acute-aged care interface: Exploring the dynamics of bed blocking. *Australasian J. Ageing* 27(3):116–120.

United Nations Population Fund (2014) News of ageing 2014. Accessed October 13, 2015, http://www.unfpa.org/ageing.

Whitt W (2002) *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues* (Springer, New York).

Whitt W (2007) What you should know about queueing models to set staffing requirements in service systems. *Naval Res. Logist.* 54(5): 476–484.

Whitt W (2013) OM forum—offered load analysis for staffing. *Manufacturing Service Oper. Management* 15(2):166–169.

Wolstenholme E (1999) A patient flow perspective of UK health services: Exploring the case for new intermediate care initiatives. *System Dynam. Rev.* 15(3):253–271.

World Health Organization (2014) Innovation for ageing populations – Addressing the challenges of frailty and disability. Accessed June 24, 2014, http://www.who.int/kobe_centre/ageing/en/.

Xie H, Chaussalet TJ, Millard PH (2005) A continuous time Markov model for the length of stay of elderly people in institutional long-term care. *J. Royal Statist. Soc. Ser. A* 168(1):51–61.

Yom-Tov GB, Mandelbaum A (2014) Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing Service Oper. Management* 16(2):283–299.

Zhang Y, Puterman ML, Nelson M, Atkins D (2012) A simulation optimization approach to long-term care capacity planning. *Oper. Res.* 60(2):249–261.

Zohar E, Mandelbaum A, Shimkin N (2002) Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Sci.* 48(4):566–583.