1. Problem definition: This research focuses on elderly patients who have been hospitalized, are ready to be discharged but 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 due to its economic implications and quality-of-life effect on patients. Indeed, hospital-delayed patients, who cannot access their most appropriate treatment (e.g., rehabilitation), prevent new admissions. Moreover, bed-blocking is costly since 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 in order to improve their joint operation.

2. Academic/Practical Relevance: The joint modeling we suggest is necessary in order to capture blocking effects. In contrast to previous research, we address an entire time-varying network by explicitly considering blocking costs. Moreover, 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.

3. 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, we analyze the fluid model and its offered-load counterpart.

4. 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. 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 demonstrates that significant cost reductions are achievable, when compared to current operations.

5. Managerial Implications: Our model can help healthcare managers in allocating geriatric beds to reduce operational costs. Moreover, we offer two new capacity allocation approaches: a periodic reallocation of beds and the incorporation of setup cost into bed allocation decisions.
1. Introduction

Providing high quality healthcare services for the ageing population is becoming a major challenge in developed countries. This challenge is amplified in view of the fact that the number of elderly people, aged 65 and over that today accounts for 10% of the population, will double within two decades (World Health Organization 2014, United Nations Population Fund 2014). In addition, the average life expectancy in nursing homes has increased from 18 months to 36 months in the last two decades (Israel Gerontological Data Center 2012).

Elderly patients, who are often frail, undergo frequent hospitalization. Indeed, about 25% of elderly admissions to Emergency Departments (EDs) are readmissions, defined as people who return within a month and require hospitalization (Tierney and Worth 1995, García-Pérez et al. 2011). The result is high occupancy levels in hospital wards and ED. For example, in the last several years, some OECD countries reported averages of over 90% occupancy levels in hospital wards (OECD iLibrary - Health at a Glance 2013, NHS England - Bed Availability and Occupancy Data 2015).

The bed blocking problem occurs when hospital patients are ready to be discharged, but must remain in the hospital until a bed in 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 since 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 UK exceeds 1.2 billion dollars per year (BBC News 2016).

In contrast to previous models, which relied on simulations for modeling bed-blocking, our research offers an analytical model for minimizing the operational costs of a system consisting of hospitals and geriatric institutions. We focus on long-term geriatric bed allocation by considering the environment described in Figure 1. It includes the community, hospitals, nursing homes and geriatric institutions. In our setting, the central decision maker is a large healthcare organization,
which operates both hospitals and geriatric institutions. In some countries (e.g., Singapore and Israel), the government functions as this organization. In England, the NHS, an arm of the government, is the central decision maker and in the U.S., it is the Veterans Administration (VA), with its 500+ hospitals, that is a major decision maker.

**Figure 1**  Patient flow network through the community, hospital wards, nursing homes and geriatric institutions

Patient flow (Figure 1) begins when elderly people turn to the ED due to a clinical deterioration or a health crisis. After stabilizing their condition, doctors decide on discharge or hospitalization. Patients can also be hospitalized without going through the ED in cases of elective procedures. Upon treatment completion, hospital doctors decide whether the patient is capable of returning to the community, needs to be admitted to a nursing home, or requires further treatment in a geriatric hospital. We subdivide the latter option into the three most common geriatric wards: rehabilitation, mechanical ventilation and skilled nursing care.
Patients who are sent to a geriatric rehabilitation ward stay there, in order to return to full
or partial functioning, one month on average. 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 5–6 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, due to 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.

We focus on four stations, which are depicted in Figure 1: Hospital wards (Station 1), Geriatric
Rehabilitation (Station 2), Mechanical Ventilation (Station 3) and Skilled Nursing Care (Station
4). Jointly analyzing these stations, for which there are long waiting lists, will lead to policies for
reducing total operational costs.

In our analysis we use two data sets. The first covers the patient flow in a hospital chain compris-
ing four general hospitals and three geriatric hospitals (three rehabilitation wards, two mechanical
ventilation wards and three skilled nursing wards). The second includes individual in-hospital wait-
ing lists for each geriatric ward. (Details about our data are provided in Appendix A.) Our data
indicates 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, it is not, 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 number of patients who are referred
to a rehabilitation ward is 5 and 9 times that of the corresponding numbers for skilled nursing care
and mechanical ventilation, respectively; this implies (Section 4.1) that the work they offer to the
system exceeds that of the other patients.
1.1. Motivation

These congestion problems, and their highly significant effect both medically and financially, motivated us to model and analyze the system depicted in Figure 1. Figure 2 presents the waiting list lengths (daily resolution) within the hospital, for each geriatric ward over one calendar year. The dotted lines represent length according to our data, while the solid lines represent our fluid model (Equations (6)–(8) in the sequel). 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  Waiting list length in hospital for each geriatric ward - model vs. data. The X axis is one calendar year in units of days.

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

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, Yom-Tov and Mandelbaum 2014, Cohen et al. 2014). Moreover, fluid models yield analytical insights, which typically cannot be obtained using other alternatives (e.g., simulation, time-varying stochastic queueing networks). We use our fluid model and its offered-load counterpart to develop and solve bed allocation problems for geriatric wards. Our goal is to find the optimal number of geriatric beds in order to minimize the total operational costs of the system.

2. Literature Review

The review covers the main areas that are relevant to this research: high-level modeling of healthcare systems, queuing 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 are 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 2–3, in order to characterize steady-state performance such as length of stay (LOS) at each station. For example, Harrison and Millard (1991) analyze the empirical distribution of patient LOS in geriatric departments by fitting a sum of two exponentials to a data set: most patients are discharged or die shortly after admission, while 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, Xie et al. 2005, Faddy and McClean 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. Differently from 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 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 demonstrate 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. While our suggested fluid model is also deterministic, we are able to justify it as the fluid limit of the corresponding stochastic model.

Discrete event simulation is another popular approach for analyzing complex systems and phenomena such as bed-blocking. El-Darzi et al. (1998) describe patient flow through geriatric wards, by examining the impact of bed-blocking and occupancy on patient flow. 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 community, hospitals and geriatric hospitals. They approximate the delay in discharge from hospital and the involved costs. Shi et al. (2015) and Armony et al. (2015) suggest a two-time-scale (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-time-scale service time captures both treatment time and additional service time caused by operational factors, such as discharge schedule.

In this research, we develop a time-varying analytical model, for setting bed capacities in geriatric institutions. Our model evolves on a single time-scale, which is days since, for the decisions we are interested in (and the data we have), it is 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 (BAS) mechanism, which happens when a patient attempts
to enter a full-capacity Station \( j \) upon completion of treatment at Station \( i \). Since 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 with 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, 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, while 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 post-operative units.

Bretthauer et al. (2011) suggest 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 time-dependent arrival patterns on the required number of operational beds in clinical wards. The authors use the offered-load approximation and the square-root staffing formula for calculating the required staffed beds for each day of the week.

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 (see Whitt 2002, Chapter 5.2 and Garnett et al. 2002). Reflection modeling, however, requires the use of indicators, which cause technical continuity problems when calculating their limits. We circumvent this challenge by developing a fluid model with blocking yet without reflection, which enables us to prove convergence of our stochastic model without reflection. Our simple and intuitive model compared to models with reflection, enables us to model 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 the 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 (MOL) approximation.

2.4. Bed Planning for Long-term Care Facilities

Most research on bed planning in healthcare systems focuses on short-term facilities, such as general hospitals (Green 2004, Akcali et al. 2006). Research about bed planning for long-term care facilities is scarce. We now review the relevant 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 estimate the waiting time for a bed in a nursing home and to find the minimal number of beds in order to achieve a target waiting time. De Vries and Beekman (1998) present a deterministic dynamic model for expressing waiting lists and waiting times of psycho-geriatric 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 United States Medicare and determine the recruiting rates of short and long stay patients to maximize profitability of the hospice. Another related research is from the telecommunication field. Jennings et al. (1997) find the optimal number of leased private lines for profit maximization. Their analysis is non-stationary, since average leased line durations are too
long (years) for the system to reach steady state within the observation period. Since hospitaliza-
tion time is long when compared to the planning horizon, transient analysis is also relevant in
the context of geriatric hospitalization. The idea of a periodic reallocation of beds to services was
suggested by Kao and Tung (1981) to minimize the expected overload at a healthcare facility. They
suggest determining the required number of beds to each service, and then for each period, transfer
one bed at a time to another service. In the suggested model, allocation decisions are based on an
analytical model and a closed-form solution for each period.

2.5. Contributions

The main contributions of our research are:

1. Modeling: We develop and analyze an analytical model comprising hospitals that interact
with geriatric wards. This joint modeling is necessary in order to capture blocking effects. The
contribution to previous research, which has focused on a single-station utility maximization (Jen-
nings et al. 1997), is in addressing an entire network analysis. This is done by explicitly considering
geriatric ward blocking costs, and minimizing the overall underage and overage cost of the system.

2. Methodology: Our work contributes to the literature on queueing (fluid) networks with block-
ing. In particular, the suggested fluid model captures blocking without the need for reflection (see
Section 2.2). Our modeling approach of blocking without reflection, applies to general networks
(for example, networks with multiple stations in tandem). We use our model to derive analytical
solutions and insights about cost minimization and bed allocation policies. The modeling approach
accommodates time-varying systems, as prevalent in healthcare, jointly with finite capacity con-
siderations, patient mortality and readmissions.

3. Practice: This research offers new capacity allocation approaches; specifically, a closed-form
solution for a periodic reallocation of beds to account for seasonal demand, and an analytical model
that incorporates setup costs into bed allocation decisions.

3. The Model

In this section, we describe our hospital environment and its dynamics. We then formally introduce
model notations and equations.
3.1. Environment, Dynamics and Notations

Consider the four framed stations in Figure 1: hospital wards (Station 1), rehabilitation (Station 2), mechanical ventilation (Station 3), and skilled nursing care (Station 4). Our model is strategic; 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 since, 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).

Table 1 summarizes the notations we use. We model the exogenous arrival rate to hospital wards as a continuous time-varying function \( \lambda(t) \) (see 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 an unlimited queue capacity, since the ED serves as a queue buffer for them (our model does, nevertheless, accommodate blocking of the first station). Patients leave the queue either when a bed becomes available or if they, unfortunately, die. Medical treatment is performed at a known service rate \( \mu_1 \). Upon treatment completion, patients are discharged back to the community, admitted to nursing homes, or referred to a geriatric ward (2, 3 or 4) with routing probabilities \( p_{1i}(t), i = 2,3,4 \), respectively. 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. Since the majority of patients are old, sick and frail,
Table 1 Notations

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda(t)$</td>
<td>External arrival rate to Station 1 at time $t$</td>
</tr>
<tr>
<td>$\lambda_i(t)$</td>
<td>Arrival rate to Station $i$ at time $t$</td>
</tr>
<tr>
<td>$\lambda_{total}(t)$</td>
<td>Total arrival (external and internal) rate to Station 1</td>
</tr>
<tr>
<td>$\mu_i$</td>
<td>Treatment rate at Station $i$</td>
</tr>
<tr>
<td>$\theta_i$</td>
<td>Individual mortality rate at Station $i$</td>
</tr>
<tr>
<td>$\beta_i$</td>
<td>Readmission rate from Station $i$ back to hospital</td>
</tr>
<tr>
<td>$N_i$</td>
<td>Number of beds at Station $i$</td>
</tr>
<tr>
<td>$x_1(t)$</td>
<td>Number of arrivals to Station 1 that have not completed their treatment at Station 1 at time $t$.</td>
</tr>
<tr>
<td>$x_i(t)$</td>
<td>Number of patients that have completed treatment at Station 1, require treatment at Station $i$, but have not yet completed their treatments at Station $i$ at time $t$</td>
</tr>
<tr>
<td>$\delta_r(t)$</td>
<td>Treatment completion rate at Station 1 at time $t$</td>
</tr>
<tr>
<td>$\delta_{total}(t)$</td>
<td>Total departure (mortality and treatment completion) rate from Station 1</td>
</tr>
<tr>
<td>$p_{ij}(t)$</td>
<td>Routing probability from Station $i$ to $j$ at time $t$</td>
</tr>
<tr>
<td>$q_i(t)$</td>
<td>Number of patients in Station $i$ at time $t$</td>
</tr>
<tr>
<td>$r_i(t)$</td>
<td>Offered load in Station $i$ at time $t$</td>
</tr>
<tr>
<td>$b_i(t)$</td>
<td>Number of blocked patients destined for Station $i$ and waiting in Station 1 at time $t$</td>
</tr>
<tr>
<td>$C_o$</td>
<td>Overage cost per day per bed</td>
</tr>
<tr>
<td>$C_u$</td>
<td>Underage cost per day per bed</td>
</tr>
</tbody>
</table>

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, refer to it as “abandonments” that can occur while waiting or while being treated. Although we use the same mortality rates while waiting and while being treated, if data prevail, our model can easily accommodate different mortality rates (see Remark 1).
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 describing the rate of change for each $q_i$. This direct approach has led to analytically intractable models that could not be justified as fluid limits of their corresponding stochastic counterparts. Moreover, these alternatives included indicator functions which, as already mentioned, are harder to analyze due to 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 deduce $q_i$ from $x_i$. This novel modeling approach also simplifies the convergence proof of the corresponding stochastic model, which is provided in Appendix B.

The value $x_1(t)$ denotes the number of arrivals to Station 1 that 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 that have completed treatment at Station 1, 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 differential equations (DEs); each characterizes the rate of change in the number of patients at each state at time $t$. The DE for $x_1$ is

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

where $\lambda_{total}(t)$ is the arrival rate to Station 1 at time $t$ and $\delta_{total}(t)$ is its departure rate.

Patients arrive to Station 1 from two sources: externally, according to rate $\lambda(t)$, and internally from Stations 2, 3 and 4. Since $\beta_i$ is the readmission rate from Station $i$ back to Station 1, the internal arrival rate to Station 1 is $\sum_{i=2}^{4} \beta_i (x_i(t) \land N_i)$, where $x \land y = \min(x,y)$; here, $(x_i(t) \land N_i)$ denotes the number of patients in treatment at Station $i$. The total arrival rate to Station 1 at time $t$ is, therefore,

$$\lambda_{total}(t) = \lambda(t) + \sum_{i=2}^{4} \beta_i (x_i(t) \land N_i). \quad (2)$$
The departure of patients who have not yet completed their service at Station 1 consists of two types, namely

$$\delta_{\text{total}}(t) = \theta_1 x_1(t) + \delta_r(t). \quad (3)$$

The first departure type is due to patients who die at an individual mortality rate $\theta_1$. Since patients might die while being hospitalized or waiting in queue, the rate at which patients die is $\theta_1 x_1(t)$.

**Remark 1.** If data regarding different mortality rates while waiting ($\theta_{1q}$) and while being treatment ($\theta_{1t}$) prevail, then the total mortality from Station 1 would be

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

The second departure type in (3), $\delta_r(t)$, is of patients who complete their treatment at Station 1. The number of blocked patients waiting in Station 1 for a transfer to Station $i$ is $(x_i(t) - N_i)^+$. Therefore, the number of unblocked beds at Station 1 is $\left( N_1 - \sum_{i=2}^{4} (x_i(t) - N_i)^+ \right)$. The rate at which patients complete their treatment in Station 1 is

$$\delta_r(t) = \mu_1 \left[ x_1(t) \wedge \left( N_1 - \sum_{i=2}^{4} (x_i(t) - N_i)^+ \right) \right], \quad (5)$$

where the expression in the rectangular brackets indicates the number of occupied unblocked beds at Station 1.

Using similar principles, we construct the DEs for the rate of change in $x_i$, $i = 2, 3, 4$. Recall that $x_i(t)$ denoted the number of patients being treated at Station $i$ and the number of blocked patients in Station 1 waiting to be transferred to Station $i$, all at time $t$. Therefore, the arrival 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,
\[
\dot{x}_1(t) = \lambda_{total}(t) - \delta_{total}(t) \\
\dot{x}_i(t) = p_{1i} \cdot \delta_i(t) - \beta_i(x_i(t) \land N_i) - \theta_i x_i(t) - \mu_i(x_i(t) \land N_i), \quad i = 2, 3, 4.
\] 

(6)

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

\[
q_1(t) = x_1(t) + \sum_{i=2}^{4} (x_i(t) - N_i)^+; \\
q_i(t) = x_i(t) \land N_i, \quad i = 2, 3, 4.
\]

(7)

Note that the number of blocked patients at Station 1 at time \(t\), waiting for an available bed at Station \(i\), \(i = 2, 3, 4\), is given by

\[
b_i(t) = (x_i(t) - N_i)^+.
\]

(8)

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

4. The Bed Allocation Model

The decision maker in our analysis is an organization that operates both general and geriatric hospitals. The objective is to find the optimal number of beds for each geriatric ward, so as to minimize overall long-term underage and overage cost 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. 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; hence, it is the per bed per day cost of hospitalization in geriatric hospitals. Underage cost, \(C_u\), is incurred when patients are delayed in the hospital due to 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; hence, $C_u$ is the per bed per day cost of hospitalization in general hospitals minus the per bed per day cost in geriatric hospitals. 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 by not providing the proper medical treatment, and by exposing the patient to diseases and contaminations prevalent in hospitals. The sum of $C_o$ and $C_u$, which will later on appear in the optimal solution in (16), includes the per bed per day hospitalization cost in general hospitals. Excluding or underestimating the cost of risk will yield a lower bound for the required number of beds. Since our solution serves as a guide for thinking, meaningful insights can be derived from such a lower bound. We denote by $C_{oi}$ and $C_{ui}$ the overage and underage costs, respectively, for Stations $i$, $i = 2, 3, 4$.

The resulting overall cost for Stations 2, 3 and 4 over a planning horizon $T$, is

$$C^{(0)}(N_2, N_3, N_4) = \sum_{i=2}^{4} C^{(0)}(N_i), \quad (9)$$

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

$$C^{(0)}(N_i) = \int_{0}^{T} \left[ C_{ui} \cdot b_i(t) + C_{oi} \cdot (N_i - q_i(t)) \right] dt, \quad i = 2, 3, 4; \quad (10)$$

here the first term in the integration is the underage cost, calculated by adding up the number of blocked patients, and the second term is the overage cost calculated via the total number of vacant beds.

Minimizing (9) is analytically intractable, since $q_i(t)$ 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 (see Jennings et al. 1997, Whitt 2007) Thus, in Sections 4.2 and 4.3 we introduce and solve, in a closed form, a problem for minimizing the total underage and overage cost based on the offered load. Then, in Section 4.4, 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 capacity are ample (no one queues up prior to service). In our context, offered-load analysis is important for understanding demand, in terms of patient-bed-days per day for the geriatric wards, in order to determine appropriate bed capacity levels.

The calculation of the offered load is carried out by solving (6) and (7) with an unlimited capacity in Stations 2, 3 and 4 \((N_i = \infty, i = 2, 3, 4)\). Note that, in this case, \( 1_{\{q_i(t) < N_i\}} \equiv 1 \) and \( b_i(t) \equiv 0 \), for \( i = 2, 3, 4 \). Under these conditions, the solutions \( q_2(t), q_3(t) \) and \( q_4(t) \) to (7) are the offered load of Stations 2, 3 and 4, respectively.

4.2. Estimating the Optimal Number of Beds based on the Offered load

The estimated overall cost for Stations 2, 3 and 4, based on the offered load over the planning horizon \( T \), is

\[
C(N_2, N_3, N_4) = \sum_{i=2}^{4} C(N_i); \tag{11}
\]

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

\[
C(N_i) = \int_0^T \left[ C_{u_i} \cdot (r_i(t) - N_i)^+ + C_{o_i} \cdot (N_i - r_i(t))^+ \right] \, dt, \quad i = 2, 3, 4. \tag{12}
\]

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 over the planning horizon. The second term, the overage cost, is obtained by multiplying \( C_{o_i} \) with the bed surplus \( (N_i - r_i(t))^+ \) and integrating. The offered load for each station is a known function of \( t \), that depends solely on input parameters but not on \( N_1, N_2, N_3 \). Thus, problem (11) is, in fact, a separable problem, which can be solved for each station separately. (When doing so below, we shall omit the \( i \) in (12) for simplicity of notations.)

To minimize \( C(N) \), we adopt the approach of Jennings et al. (1997) and treat \( N \) as a continuous variable. We let \( r_d = \{r_d(t)\} \) denote the *decreasing rearrangement* of \( r \) on the interval [0, \( T \]).


[0, T]; \( r_d \) on \([0, T]\) is characterized by being the unique decreasing function such that, for all \( x \geq 0 \), we have

\[
\int_0^T 1_{\{r(t) \geq x\}} dt = \int_0^T 1_{\{r_d(t) \geq x\}} dt; \tag{13}
\]

here \( 1_{\{r(t) \geq x\}} \) denotes the indicator function for the event \( \{r(t) \geq x\} \). Existence and uniqueness of \( r_d \) were established in Hardy et al. (1952). The interpretation of Equation (13) 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:

\[
C(N) = \int_0^T [C_u \cdot (r(t) - N)^+ + C_o \cdot (N - r(t))^+] dt
\]

\[
= \int_N^\infty C_u \int_0^T 1_{\{r(t) \geq x\}} dx + \int_0^N C_o \int_0^T 1_{\{r(t) \leq x\}} dx
\]

\[
= \int_N^\infty C_u \int_0^T 1_{\{r(t) \geq x\}} dx - \int_0^N C_u \int_0^T 1_{\{r(t) \geq x\}} dx + \int_0^N C_o [T - \int_0^T 1_{\{r(t) \geq x\}}] dx
\]

\[
= \int_N^\infty C_u \int_0^T 1_{\{r(t) \geq x\}} dx - \int_0^N (C_u + C_o) \int_0^T 1_{\{r(t) \geq x\}} dx + C_o TN
\]

\[
= \int_N^\infty C_u \int_0^T 1_{\{r_d(t) \geq x\}} dx - \int_0^N (C_u + C_o) \int_0^T 1_{\{r_d(t) \geq x\}} dx + C_o TN,
\]

(14)

where the first equality is achieved by substituting:

\[
(r(t) - N)^+ = \int_N^\infty 1_{\{r(t) \geq x\}} dx, \quad (N - r(t))^+ = \int_0^N 1_{\{r(t) \leq x\}} dx,
\]

(15)

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 Appendix C. 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

\[
N^* = r_d \left( \frac{C_o T}{C_o + C_u} \right).
\]
Note that $C_o + C_u$ is actually the per day per bed cost in general hospitals. Hence, the ratio $C_o/(C_o + C_u)$ in the optimal solution is the hospitalization cost ratio for a geriatric bed and a general hospital bed. In Appendix D we explain how $N^*$ arose as a candidate for minimizing $C(N)$.

**Remark 2.** 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. However, we believe that our solution approach is more intuitive for the present problem. Moreover, it naturally enables the solution of an extension which, includes setup cost per new bed, as discussed in Section 5.2.

### 4.3. An Illustrative Example

In this section, we apply our model to data, in order to calculate the optimal number of beds in each geriatric ward. We then compare our solution to the current situation.

The experts we consulted with estimated that underage cost is about four times the overage cost ($C_u \approx 4C_o$). This is because hospitalization costs in general hospitals are much higher than in a geriatric hospital. The same ratio between $C_u$ and $C_o$ was estimated for all three geriatric wards; however, accommodating a different ratio to each ward can be easily implemented. We used the fluid model developed in Section 3, together with historical data, to forecast the offered load for a three-year planning horizon, where the demand for beds (e.g., the arrival rate) increases every year. Then, by using Matlab we numerically calculated the functions $r_d$ for each ward. The optimal number of beds is the value of these functions at the critical point as in (16). Since the value of $N^*$ is not necessarily an integer, it must be rounded. Rounding up vs. down has minor significance, since 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 presents the optimal number of beds (the dashed lines) compared to the offered load (continuous lines). The optimal number of beds for each ward was calculated by rounding up the result from Equation (16). The optimal solution implies
increasing the number of geriatric beds by 37%. This will lead to an overage and underage cost reduction of about 78%, when compared to the cost under the current number of beds for the same arrival forecast. A similar cost reduction (about 74%) is achieved when comparing the current and optimal number of beds using our corresponding simulation model of the system. Details regarding this simulation model are provided in Appendix A. We believe there are two major reasons for this dramatic cost reduction. The first is due to difficulties in increasing the present budget for the introduction of new beds. The second reason is the lack of a model, as the one introduced here, that takes blocking and its related costs into account, and which will guide planners. We provide more details and calculate imputed costs in Section 4.5.

Figure 3  Optimal number of beds

Suppose, for comparison sake, that there is flexibility between the three geriatric wards. In other words, a specific geriatric bed can be allocated to another ward if needed. Then, we could solve Equation (11) for the total offered load. The optimal solution will then require fewer beds overall (602 beds instead of 613) but will lead to only an additional decrease of 5% in the total cost. The reason for this relatively modest advantage is the similar offered-load patterns among the wards.
4.4. Cost Comparison

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 (9), is based on the time-varying number of patients, as derived from the solution of the fluid equations in (7). Since minimizing \( C^{(0)}(N_2, N_3, N_4) \) is analytically intractable, we introduced the second cost function, \( C(N_2, N_3, N_4) \) in (11), which estimates the total cost based on an offered-load approximation to the time-varying demand for beds.

In order 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. All parameters, including the size of the system, are realistic for the system we analyze. The solution for \( C^{(0)}(N_2, N_3, N_4) \) was calculated by our closed-form expression in (16). The solution for \( C(N_2, N_3, N_4) \) was achieved by numerically solving the optimization problem; this was done by choosing the capacity combination with the minimal cost. Finally, the solution for the simulation model was achieved by calculating, for each capacity combination, the total underage and overage cost. This was done by using (9) and (10), 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 which minimized the cost. Table 2 summarizes this comparison by comparing the optimal number of beds with the optimal cost for each method. The maximal difference between the solutions is 1.1%, when comparing bed allocations and 3.6% 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.

4.5. The Imputed Overage and Underage Costs

In addition to the estimation of the \( C_o/C_u \) ratio given to us by experts, 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 number of beds that decision makers allocate to each geriatric ward. To this end, we use the current number of beds in each geriatric ward in order to extract the model’s parameters.
Table 2  Comparing optimal solutions (number of beds and cost per year) \(-C(0)(N_2,N_3,N_4)\) vs. \(C(N_2,N_3,N_4)\) vs. simulation.

<table>
<thead>
<tr>
<th>Ward</th>
<th>(C(0)(N_2,N_3,N_4))</th>
<th>(C(N_2,N_3,N_4))</th>
<th>Simulation</th>
<th>Maximal difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(N^*) (Total cost)</td>
<td>(N^*) (Total cost)</td>
<td>(N^*) (Total cost)</td>
<td>(N^*) (Total cost)</td>
</tr>
<tr>
<td>Rehabilitation</td>
<td>306 (2,480,333)</td>
<td>303 (2,616,250)</td>
<td>305 (2,548,000)</td>
<td>1.0% (5.2%)</td>
</tr>
<tr>
<td>Mechanical Ventilation</td>
<td>153 (1,710,917)</td>
<td>151 (1,772,750)</td>
<td>153 (1,1743,583)</td>
<td>1.3% (3.5%)</td>
</tr>
<tr>
<td>Skilled Nursing</td>
<td>161 (1,117,083)</td>
<td>159 (1,117,667)</td>
<td>160 (1,115,333)</td>
<td>1.3% (0.2%)</td>
</tr>
<tr>
<td>Total Number of beds</td>
<td>620 (5,308,333)</td>
<td>613 (5,506,667)</td>
<td>618 (5,406,917)</td>
<td>1.1% (3.6%)</td>
</tr>
</tbody>
</table>

\(C_o\) 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. Consequently, from Theorem 1 we have

\[
I = \frac{r_d^{-1}(N)}{T} = \frac{C_o}{C_o + C_u}, \tag{17}
\]

where we define

\[
r_d^{-1}(N) \equiv \sup \{t| r_d(t) \geq N\}, \tag{18}
\]

as the time during which underage costs were incurred. Therefore, \(r_d^{-1}(N)/T\) is the fraction of time during which underage costs were incurred.

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:

\[
\bar{I} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{r(t_i) \geq N\}}. \tag{19}
\]

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

\[
\bar{I} = \frac{C_o}{C_o + C_u}. \tag{20}
\]

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 \(C_u = 0.35C_o\) (\(C_o \approx 3C_u\)) in rehabilitation,
\( C_u = 0.099C_o \ (C_o = 10C_u) \) in skilled nursing care and \( C_u = 0 \) in mechanical ventilation. This is opposed to \( C_u = 4C_o \), which was the experts’ estimation.

These imputed costs may imply that blocking costs are underestimated when determining the geriatric bed capacity. Another possible explanation is that although there is a central decision maker that owns both the general and geriatric hospitals, decisions are still optimized locally.

5. Extensions

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

5.1. Periodic Reallocation of Beds

Managers of geriatric institutions acknowledge that it is feasible to change the number of beds during the year in order 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 then determine the optimal number of beds for each period. For example, a semi-annual reallocation would determine a certain capacity during the first half year of every year in the planning horizon and a possibly different capacity during the second half year. The described policy is feasible since most of the ‘bed cost’ is related to labor costs and medical supplies: the latter can be purchased seasonally while the former can be changed due to 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).

A much easier problem to solve would be to find the optimal number of beds for each half year during the planning horizon (e.g., six capacity levels for a thee-year planning horizon). However, the problem we solve, which is more relevant to the healthcare environment we model, determines only two capacity levels—one for the first half of every year and one for the second half.

In the following examples, each year is divided into equal periods; nevertheless, a different division can be easily implemented to better fit the demand pattern.
5.1.1. Semi-Annual Reallocation

Let $N_1^S$ and $N_2^S$ denote the number of beds within a geriatric ward for the first and second halves of every year. Our objective is to find $N_1^S$ and $N_2^S$ that minimize the total underage and overage costs. Figure 4 illustrates such a periodic reallocation by demonstrating it on the rehabilitation ward data. The dark areas represent overage periods and the light areas represent underage periods. Here, $N_1^S = 310$ for the first half year and $N_2^S = 260$ for the second half year.

In order to find $N_1^S$ and $N_2^S$ that minimize the total cost, we separate the solution for each of the halves. We split $r(t)$ into two functions: $r_1(t)$ for the first half of every year and $r_2(t)$ for the second half. Each of the functions is defined on the interval $[0, T/2]$, by concatenating the relevant intervals from $r(t)$ and shifting the function to $t = 0$. We define functions $r_{d_1}(t)$ and $r_{d_2}(t)$ to be the decreasing rearrangement of $r_1(t)$ and $r_2(t)$, respectively, exactly as we defined $r_d(t)$ in Section 4. The total underage and overage costs are, therefore,

$$C(N_1^S, N_2^S) = C(N_1^S) + C(N_2^S)$$

$$= \int_0^{T/2} [C_u(r_1(t) - N_1^S)^+ + C_o(N_1^S - r_1(t))^+] dt$$

$$+ \int_0^{T/2} [C_u(r_2(t) - N_2^S)^+ + C_o(N_2^S - r_2(t))^+] dt. \quad (21)$$
Proposition 1. The number of beds that minimizes $C(N_1^S, N_2^S)$ is

$$
N_1^{S*} = r_{d_1} \left( \frac{C_o}{C_o + C_u} \cdot \frac{T}{2} \right), \quad N_2^{S*} = r_{d_2} \left( \frac{C_o}{C_o + C_u} \cdot \frac{T}{2} \right).
$$

(22)

Proof: Since $r_1(t)$ and $r_2(t)$ are non-negative and measurable on the interval $[0, T/2]$ (see Hardy et al. 1952), we can implement the results from Theorem 1 by replacing $T$ with $T/2$. □

The same methodology can be used to find the optimal number of beds for a quarterly reallocation. In fact, management can set the interval length for each case according to demand patterns and feasibility of reallocation points.

5.1.2. A Numerical Example

We now solve the periodic reallocation problem for semi-annual and quarterly reallocations, for a three-year planning horizon. Figure 5 depicts the semi-annual solution. The continuous lines represent the offered load for each ward, while the dashed lines represent the optimal number of beds.

Figure 5  Optimal semi-annual reallocation of beds
Table 3 shows the optimal number of beds for each periodic reallocation as well as the optimal constant number of beds. Table 4 compares the overage and underage cost reduction of each allocation policy compared to the constant bed capacity policy. The major cost reduction (by almost 70%), compared to 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—for example, quarterly reallocation policies reduce costs, on average, by almost 12%, when compared to the policy with a constant number of beds.

<table>
<thead>
<tr>
<th>Ward</th>
<th>Constant</th>
<th>Periodically</th>
<th>Periodically</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$N^*$</td>
<td>$N_s^*$</td>
<td>$N_s^*$</td>
</tr>
<tr>
<td>Rehabilitation</td>
<td>303</td>
<td>323</td>
<td>265</td>
</tr>
<tr>
<td>Mechanical Ventilation</td>
<td>151</td>
<td>160</td>
<td>138</td>
</tr>
<tr>
<td>Skilled Nursing</td>
<td>159</td>
<td>162</td>
<td>142</td>
</tr>
<tr>
<td>Total</td>
<td>613</td>
<td>645</td>
<td>545</td>
</tr>
</tbody>
</table>

Table 4 Overage and underage cost reduction, in percentages, compared to the constant bed capacity policy

<table>
<thead>
<tr>
<th>Ward</th>
<th>Semi-annual reallocation</th>
<th>Quarterly reallocation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rehabilitation</td>
<td>12%</td>
<td>15%</td>
</tr>
<tr>
<td>Mechanical Ventilation</td>
<td>3.75%</td>
<td>12.5%</td>
</tr>
<tr>
<td>Skilled Nursing</td>
<td>7.5%</td>
<td>7.5%</td>
</tr>
<tr>
<td>Total</td>
<td>7.7%</td>
<td>11.5%</td>
</tr>
</tbody>
</table>

5.2. Including Setup Cost per New Bed

In this section, we analyze a situation where a fixed setup cost, $K$, is 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 can vary across bed types.

The overall cost for a geriatric ward is then

\[ C_K(N) = C(N) + K(N - B)^+, \]  

where \( C(N) \) is the overall cost, analyzed in Section 4; \( B \) is the current bed capacity 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

\[
N^*_K = \begin{cases} 
  r_d \left( \frac{C_o T}{C_o + C_u} \right), & \text{if } r_d \left( \frac{C_o T}{C_o + C_u} \right) \leq B \\
  r_d \left( \frac{C_o T + K}{C_o + C_u} \right), & \text{if } r_d \left( \frac{C_o T + K}{C_o + C_u} \right) \geq B \\
  B, & \text{otherwise.}
\end{cases}
\]  

We prove Theorem 2 in Appendix E.

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

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

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 when there is a high bed setup cost it is preferable to pay the underage cost for the entire planning horizon.

Note that the optimal solution depends on the available bed capacity. For a very large \( B \), there is no point introducing new beds and, therefore, the optimal solution equals the solution with no setup cost. On the other hand, if the current capacity, \( B \), is very small, then adding new beds is essential for decreasing the total cost. In all other cases, it may be preferable to keep the capacity as is.
6. Future Research

There are several directions worthy of future research. The first is to modify the structure of the system by adding an intermediate ward (i.e., a step-down unit) for sub-acute geriatrics, between the general hospital and the geriatric hospitals. Such an intermediate ward would be designated for elderly patients with an expected long stay in the general hospital, before continuing on to a geriatric ward. Adding a sub-acute ward can both reduce the workload and bed occupancy in hospitals and improve the patient flow in and out of the hospitals.

The second direction is to include home healthcare services or virtual hospitals (Ticona and Schulman 2016), that could be modeled as another type of bed using our framework, with different costs and treatment durations. Another direction is a capacity allocation analysis where given a predefined budget, the decision maker must decide where it is most beneficial to add new beds: in general hospitals, in intermediate wards or in the geriatric wards. The simple version of this question (without intermediate wards), in fact, triggered the present research.

References


Liu, Y., W. Whitt. 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. 37


Whitt, W. 2007. What you should know about queueing models to set staffing requirements in service systems. *Naval Research Logistics (NRL)* **54**(5) 476–484. 9, 16


Appendix A: Model Validation

To validate our model we used the following patient flow data:

1. Two years of patient flow data from a district that includes four general hospitals and three geriatric hospitals (three rehabilitation wards, two mechanical ventilation wards and three skilled nursing wards).

2. Two years of waiting lists for geriatric wards, including individual waiting times from our Partner Hospital.

Based on the patient flow data, model parameters were first estimated, then inspected and validated by expert doctors. The parameter values used for the validation are: $\mu_1 = 0.85$, $\mu_2 = 0.3$, $\mu_3 = 0.16$, $\mu_4 = 0.04$, $\beta_2 = 0.02$, $\beta_3 = 0.01$, $\beta_4 = 0.01$, $\theta_1 = 125$, $\theta_2 = 0.0025$, $\theta_3 = 0.001$, $\theta_4 = 0.001$, $N_1 = 600$, $N_2 = 226$, $N_3 = 93$, $N_4 = 120$ (we used day as a time unit). For example, Station 1 contains 600 beds; the average treatment duration there is 4.85 days and the average time to death is 125 days.

Estimating the rates of mortality and readmission were done using the MLE (Maximum Likelihood Estimator), that is prevalent for estimating censored data, such as patience and retries in service systems (see Zohar et al. 2002 for details). Here, we adjust the estimator for the case where patients die while being in treatment, rather than just while waiting in queue. To this end, instead of the actual waiting time, we consider the actual treatment time.

The time-varying arrival rates and routing probabilities were also derived from the data. The average monthly arrival rate was 3,632 patients per month (with a minimum of 3,559 and maximum 3,774), and the average routing probabilities to each geriatric ward were 9% for rehabilitation wards, 0.8% for mechanical ventilation and 2.4% for skilled nursing care.

Using these parameters, we numerically (via Matlab) solved (7), which resulted in the number of patients in each ward at any time ($q_i(t)$ for $i = 1, 2, 3, 4$) and the number of blocked patients waiting for each ward ($b_i(t)$ for $i = 2, 3, 4$). Figure 2 shows the length of the waiting lists for each ward, using a daily resolution during one calendar year, according to the data and the fluid model. The very good fit implies that the fluid model is appropriate for modeling the system considered here. The three geriatric wards work at full capacity throughout the year; there are always blocked patients in the hospital and any vacant geriatric bed is immediately filled.

In addition to comparing the fluid model with real data, we validated its accuracy against a discrete event simulation of a stochastic system, which we developed for this purpose in Mathworks$^TM$ – SimEvents/MATLAB. We conducted experiments for several scenarios; in each one, we considered three levels of the scaling parameter $\eta$. In our simulation model, the patients arrive according to a non-homogeneous Poisson process that was used to represent general, time-dependent arrivals, as prevalent in hospitals (Bekker and de Bruin 2010, Yom-Tov and Mandelbaum 2014, Shi et al. 2015, Armony et al. 2015). The treatment rates were randomly generated from exponential, Phase-type (as a mixture of two exponentials) and Lognormal distributions, which are typical for describing lengths of stay in hospitals and geriatric wards (McClean and Millard 1993, Marazzi et al. 1998, Xie et al. 2005, McClean and Millard 2006, Faddy et al. 2009, Armony et al. 2015). The expectations of these three distributions were equal when compared in a specific scenario.
For each scenario and $\eta$ we used 300 replications, each for 1000 days, and calculated the Root Mean Square Error (RMSE) using the following formula:

$$RMSE = \sqrt{\frac{1}{T} \sum_{t=0}^{T} \left[ q_{i,\text{sim}}(t) - q_{i,\text{fluid}}(t) \right]^2 dt};$$

(26)

Here $q_{i,\text{sim}}(t)$ is the total number of patients in Station $i$ at time $t$ according to the simulation results and $q_{i,\text{fluid}}(t)$ is the number according to the fluid model. The results are summarized in Tables 5 and 6. An example for Scenario 1 with $\eta = 10$ is illustrated in Figure 6. As expected, fluid models become more accurate as the scaling parameter $\eta$ becomes larger. In general, the best results were achieved for the Exponential distributions. However, the model is quite accurate even for the Phase-type and Lognormal distribution. In all cases, the fluid model accurately forecasts, within a 95% confidence interval, the stochastic behavior of the corresponding simulation. The percentage of error, relative to system capacity, varied from 0.6% to 2.4%. However, for the size of systems we are interested in (Scenarios 1-18), the percentage of error was less than 1%.

Figure 6  Scenario 1 in Table 5. On the right: Total number of patients in each geriatric ward - fluid model vs. simulation. On the left: The arrival rate $\lambda(t)$.

Appendix B: Fluid Model for Blocking: Convergence of the Stochastic Model

We now develop a fluid model with blocking for a network with $k$ stations, as illustrated in Figure 7. We focus on modeling the blocking and not the mortality and readmissions, which are also included in our model, since the latter two phenomena can be addressed with known techniques (see Mandelbaum et al. 1998, 1999 and Cohen et al. 2014 for details).

Our system is characterized by the following (deterministic) parameters:

1. Arrival rate to Station 1 is $\lambda(t)$, $t \geq 0$;
2. Service rate $\mu_i > 0$, $i = 1, \ldots, k$;
3. Number of servers $N_i$, $i = 1, \ldots, k$;
Table 5  Parameters of scenarios

<table>
<thead>
<tr>
<th>No.</th>
<th>$N_1, N_2, N_3, N_4$</th>
<th>$\mu_1, \mu_2, \mu_3, \mu_4$</th>
<th>$p_{12}, p_{13}, p_{14}$</th>
<th>distribution</th>
<th>$\lambda(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>600, 234, 93, 120</td>
<td>1/4.85, 1/30, 1/160, 1/45</td>
<td>0.09, 0.008, 0.024</td>
<td>Exponential</td>
<td>polynomially</td>
</tr>
<tr>
<td>2</td>
<td>600, 234, 93, 120</td>
<td>1/4.85, 1/30, 1/160, 1/45</td>
<td>0.09, 0.008, 0.024</td>
<td>Phase-Type</td>
<td>polynomially</td>
</tr>
<tr>
<td>3</td>
<td>600, 234, 93, 120</td>
<td>1/4.85, 1/30, 1/160, 1/45</td>
<td>0.09, 0.008, 0.024</td>
<td>Exponential</td>
<td>polynomially/10</td>
</tr>
<tr>
<td>4</td>
<td>600, 234, 93, 120</td>
<td>1/4.85, 1/30, 1/160, 1/45</td>
<td>0.09, 0.008, 0.024</td>
<td>Exponential</td>
<td>polynomially</td>
</tr>
<tr>
<td>5</td>
<td>600, 234, 93, 120</td>
<td>1/4.85, 1/30, 1/160, 1/45</td>
<td>0.09, 0.008, 0.024</td>
<td>Exponential</td>
<td>polynomially</td>
</tr>
<tr>
<td>6</td>
<td>600, 234, 93, 120</td>
<td>1/4.85, 1/30, 1/160, 1/45</td>
<td>0.09, 0.008, 0.024</td>
<td>Exponential</td>
<td>polynomially</td>
</tr>
<tr>
<td>7</td>
<td>600, 234, 93, 120</td>
<td>1/4.85, 1/30, 1/160, 1/45</td>
<td>0.09, 0.008, 0.024</td>
<td>Exponential</td>
<td>polynomially</td>
</tr>
<tr>
<td>8</td>
<td>600, 234, 93, 120</td>
<td>1/4.85, 1/30, 1/160, 1/45</td>
<td>0.09, 0.008, 0.024</td>
<td>Exponential</td>
<td>polynomially/10</td>
</tr>
<tr>
<td>9</td>
<td>600, 234, 93, 120</td>
<td>1/4.85, 1/30, 1/160, 1/45</td>
<td>0.09, 0.008, 0.024</td>
<td>Exponential</td>
<td>polynomially</td>
</tr>
<tr>
<td>10</td>
<td>600, 200, 200, 200</td>
<td>1/5, 1/30, 1/30, 1/30</td>
<td>0.25, 0.25, 0.25</td>
<td>Exponential</td>
<td>polynomially</td>
</tr>
<tr>
<td>11</td>
<td>600, 200, 200, 200</td>
<td>1/5, 1/30, 1/30, 1/30</td>
<td>0.25, 0.25, 0.25</td>
<td>Exponential</td>
<td>polynomially</td>
</tr>
<tr>
<td>12</td>
<td>600, 200, 200, 200</td>
<td>1/5, 1/30, 1/30, 1/30</td>
<td>0.25, 0.25, 0.25</td>
<td>Exponential</td>
<td>polynomially</td>
</tr>
<tr>
<td>13</td>
<td>600, 200, 200, 200</td>
<td>1/5, 1/30, 1/30, 1/30</td>
<td>0.25, 0.25, 0.25</td>
<td>Exponential</td>
<td>polynomially</td>
</tr>
<tr>
<td>14</td>
<td>600, 200, 200, 200</td>
<td>1/5, 1/30, 1/30, 1/30</td>
<td>0.25, 0.25, 0.25</td>
<td>Exponential</td>
<td>polynomially</td>
</tr>
<tr>
<td>15</td>
<td>600, 200, 200, 200</td>
<td>1/5, 1/30, 1/30, 1/30</td>
<td>0.25, 0.25, 0.25</td>
<td>Exponential</td>
<td>polynomially</td>
</tr>
<tr>
<td>16</td>
<td>600, 200, 200, 200</td>
<td>1/5, 1/30, 1/30, 1/30</td>
<td>0.25, 0.25, 0.25</td>
<td>Exponential</td>
<td>polynomially</td>
</tr>
<tr>
<td>17</td>
<td>600, 200, 200, 200</td>
<td>1/5, 1/30, 1/30, 1/30</td>
<td>0.25, 0.25, 0.25</td>
<td>Exponential</td>
<td>polynomially</td>
</tr>
<tr>
<td>18</td>
<td>600, 200, 200, 200</td>
<td>1/5, 1/30, 1/30, 1/30</td>
<td>0.25, 0.25, 0.25</td>
<td>Exponential</td>
<td>polynomially</td>
</tr>
<tr>
<td>19</td>
<td>60, 20, 20, 20</td>
<td>1/5, 1/30, 1/30, 1/30</td>
<td>0.09, 0.008, 0.024</td>
<td>Exponential</td>
<td>polynomially/10</td>
</tr>
<tr>
<td>20</td>
<td>60, 20, 20, 20</td>
<td>1/5, 1/30, 1/30, 1/30</td>
<td>0.09, 0.008, 0.024</td>
<td>Exponential</td>
<td>polynomially/10</td>
</tr>
<tr>
<td>21</td>
<td>60, 20, 20, 20</td>
<td>1/5, 1/30, 1/30, 1/30</td>
<td>0.09, 0.008, 0.024</td>
<td>Exponential</td>
<td>polynomially</td>
</tr>
</tbody>
</table>

Table 6  Total number in each station - fluid model vs. Simulation - RMSE results

<table>
<thead>
<tr>
<th>No.</th>
<th>$\eta = 1$</th>
<th>$\eta = 10$</th>
<th>$\eta = 100$</th>
<th>No.</th>
<th>$\eta = 1$</th>
<th>$\eta = 10$</th>
<th>$\eta = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.07</td>
<td>2.42</td>
<td>0.89</td>
<td>12</td>
<td>11.4</td>
<td>5.12</td>
<td>1.14</td>
</tr>
<tr>
<td>2</td>
<td>8.92</td>
<td>3.52</td>
<td>1.24</td>
<td>13</td>
<td>7.42</td>
<td>2.13</td>
<td>0.78</td>
</tr>
<tr>
<td>3</td>
<td>11.68</td>
<td>5.21</td>
<td>1.32</td>
<td>14</td>
<td>7.74</td>
<td>3.41</td>
<td>0.96</td>
</tr>
<tr>
<td>4</td>
<td>9.87</td>
<td>2.78</td>
<td>0.97</td>
<td>15</td>
<td>10.98</td>
<td>4.64</td>
<td>1.01</td>
</tr>
<tr>
<td>5</td>
<td>10.76</td>
<td>3.85</td>
<td>1.45</td>
<td>16</td>
<td>8.01</td>
<td>2.23</td>
<td>0.72</td>
</tr>
<tr>
<td>6</td>
<td>12.52</td>
<td>5.69</td>
<td>1.38</td>
<td>17</td>
<td>8.59</td>
<td>3.41</td>
<td>0.91</td>
</tr>
<tr>
<td>7</td>
<td>7.67</td>
<td>2.28</td>
<td>0.82</td>
<td>18</td>
<td>11.23</td>
<td>4.76</td>
<td>0.98</td>
</tr>
<tr>
<td>8</td>
<td>8.32</td>
<td>3.44</td>
<td>1.05</td>
<td>19</td>
<td>2.35</td>
<td>1.95</td>
<td>0.58</td>
</tr>
<tr>
<td>9</td>
<td>11.21</td>
<td>5.09</td>
<td>1.14</td>
<td>20</td>
<td>2.76</td>
<td>2.28</td>
<td>1.24</td>
</tr>
<tr>
<td>10</td>
<td>8.03</td>
<td>2.28</td>
<td>0.82</td>
<td>21</td>
<td>2.91</td>
<td>2.43</td>
<td>1.32</td>
</tr>
<tr>
<td>11</td>
<td>8.65</td>
<td>3.5</td>
<td>1.05</td>
<td></td>
<td>Avg</td>
<td>8.53</td>
<td>3.56</td>
</tr>
</tbody>
</table>

4. Transfer probability $p_{ij}(t)$ from Station $i$ to Station $j$;
5. Unlimited waiting room in Station 1;
6. No waiting room in Stations $i = 1, \ldots, k$.

To simplify notation, for the purpose of this proof we assume $p_{ij}(t) = p_{ij}$. Generalizing to time-varying transition probabilities is straightforward.
The stochastic model is created from the following stochastic building blocks $A, S_i, i = 1, \ldots, (2k - 1)$, which are assumed to be independent, as well as $X_i(0), i = 1, \ldots, k$: \[X_i(0) = \begin{cases} a_i & \text{if } z \in (0, a_i) \\ b_i & \text{if } z \in (a_i, b_i) \\ c_i & \text{if } z \in (b_i, c_i) \end{cases}\] where $a_i, b_i, c_i$ are parameters, and $z$ is the time variable.
1. External arrival process $A = \{A(t), t \geq 0\}$: $A$ is a counting process, in which $A(t)$ represents the external cumulative number of arrivals up to time $t$. The arrival rate $\lambda(t)$, $t \geq 0$ is related to $A$ via
\[
EA(t) = \int_0^t \lambda(u)du, \quad t \geq 0.
\] (27)

A special case is the non-homogeneous Poisson process, for which
\[
A(t) = A_0^t \left( \int_0^u \lambda(u)du \right), \quad t \geq 0,
\] (28)
where $A_0(\cdot)$ is a standard Poisson process (constant arrival rate 1).

2. “Basic” nominal service processes $S_i = \{S_i(t), t \geq 0\}$, $i = 1, \ldots, (2k - 1)$, where $S_i(t)$ are standard Poisson processes.

3. Initial number of customers in each state $X_i(0)$, $i = 1, \ldots, k$.

The above building blocks will yield the following $k$ stochastic process, which captures the state of our system:

The stochastic process $X_1 = \{X_1(t), t \geq 0\}$ denotes the number of arrivals to Station 1 that have not completed their service at Station 1 at time $t$.

The stochastic process $X_i = \{X_i(t), t \geq 0\}$, $i = 2, \ldots, k$ denotes the number of customers that have completed service at Station 1, require service at Station $i$, but have not yet completed their service at Station $i$ at time $t$.

We assume that the blocking mechanism is blocking-after-service (BAS) (Balsamo et al. 2001). Thus, if upon service completion at Station 1, the destination station is saturated, the customer will be forced to stay in Station 1, while occupying a server there until the destination station becomes available. The latter means that when a server completes service, the blocked customer immediately transfers and starts service.

Let $Q = \{Q_1(t), Q_2(t), \ldots, Q_k(t), t \geq 0\}$ denote the stochastic queueing process in which $Q_i(t)$ represents the number of customers at Station $i$ at time $t$. The process $Q$ is characterized by the following equations:

\[
Q_1(t) = X_1(t) + \sum_{i=2}^k (X_i(t) - N_i)^+; \\
Q_j(t) = X_j(t) \land N_j, \quad j = 2, \ldots, k;
\] (29)
here
\[
X_1(t) = X_1(0) + A(t) - \sum_{m=2}^{k} S_m \left( p_{1m} \cdot \mu_1 \int_0^t \left[ X_1(u) \wedge \left( N_1 - \sum_{i=2}^{k} (X_i(u) - N_i)^+ \right) \right] \, du \right) \\
- S_1 \left( (1 - \sum_{i=2}^{k} p_{1i}) \mu_1 \int_0^t \left[ X_1(u) \wedge \left( N_1 - \sum_{i=2}^{k} (X_i(u) - N_i)^+ \right) \right] \, du \right),
\]
\[X_j(t) = X_j(0) + S_1 \left( p_{1j} \cdot \mu_1 \int_0^t \left[ X_1(u) \wedge \left( N_1 - \sum_{i=2}^{k} (X_i(u) - N_i)^+ \right) \right] \, du \right) \\
- S_{k-1+j} \left( \mu_j \int_0^t (X_j(u) \wedge N_j) \, du \right), \quad j = 2, \ldots, k.
\]

An inductive construction over time shows that (30) uniquely determines the process X.

Note that \((X_i(t) - N_i)^+, i = 2, \ldots, k\), is the number of blocked customers waiting for an available server in Station \(i\).

### B.1. Fluid Approximation - FSLLN

We now develop a fluid limit for our queueing model through a Functional Strong Law of Large Numbers (FSLLN). We begin with (30) and scale up the arrival rate and the number of servers by \(\eta > 0, \eta \to \infty\). This \(\eta\) will serve as an index of a corresponding queueing process \(X^\eta\):

\[
X_1^\eta(t) = X_1^\eta(0) + A^\eta(t) - \sum_{m=2}^{k} S_m \left( p_{1m} \cdot \mu_1 \int_0^t \left[ X_1^\eta(u) \wedge \left( \eta N_1 - \sum_{i=2}^{k} (X_i^\eta(t) - \eta N_i)^+ \right) \right] \, du \right) \\
- S_1 \left( (1 - \sum_{i=2}^{k} p_{1i}) \mu_1 \int_0^t \left[ X_1^\eta(u) \wedge \left( \eta N_1 - \sum_{i=2}^{k} (X_i^\eta(t) - \eta N_i)^+ \right) \right] \, du \right),
\]
\[X_j^\eta(t) = X_j^\eta(0) + S_1 \left( p_{1j} \cdot \mu_1 \int_0^t \left[ X_1^\eta(u) \wedge \left( \eta N_1 - \sum_{i=2}^{k} (X_i^\eta(t) - \eta N_i)^+ \right) \right] \, du \right) \\
- S_{k-1+j} \left( \mu_j \int_0^t (X_j^\eta(u) \wedge \eta N_j) \, du \right), \quad j = 2, 3, \ldots, k.
\]

Suppose that \(A^\eta, \eta > 0\), the family of arrival processes satisfies the following FSLLN:

\[
\lim_{\eta \to \infty} \frac{1}{\eta} A^\eta(t) = \int_0^t \lambda(u) \, du;
\]
here the convergence is uniformly on compact sets of \(t \geq 0\) (u.o.c.). For example, in the non-homogeneous Poisson process

\[
A^\eta(t) = A_0 \left( \int_0^t \eta \lambda(u) \, du \right), \quad t \geq 0.
\]

Other examples can be found in Liu and Whitt (2011a, 2012, 2014).

Assumption (32) is all that is required in order to apply Theorem 2.2 in Mandelbaum et al. (1998) and get

\[
\lim_{\eta \to \infty} \frac{1}{\eta} X_i^\eta(t) = x_i(t), \quad \text{uoc}, \quad i = 1, \ldots, k,
\]

where \(x_i, i = 1, 2, \ldots, k\), are referred to as the fluid limit associated with the queueing family \(X_i^\eta, i = 1, \ldots, k\).

The functions \(x_i\) constitute the unique solution of the following ODE:

\[
x_1(t) = x_1(0) + \int_0^t \left[ \lambda(u) - \mu_1 \left( x_1(u) \wedge (N_1 - \sum_{i=2}^{k} (x_i(u) - N_i)^+) \right) \right] \, du,
\]
\[x_j(t) = x_j(0) + p_{1j} \cdot \mu_1 \int_0^t \left[ x_1(u) \wedge (N_1 - \sum_{i=2}^{k} (x_i(u) - N_i)^+) - \mu_j (x_j(u) \wedge N_j) \right] \, du, \quad j = 2, \ldots, k.
\]
We now introduce the functions $q_i, i = 1, \ldots, k,$ as the fluid limit associated with the queueing family $Q^\eta$; these functions are given by
\begin{align}
q_1(t) &= x_1(t) + \sum_{i=2}^{k} (x_i(t) - N_i)^+, \\
q_j(t) &= x_j(t) \wedge N_j, \quad j = 2, \ldots, k.
\end{align}
(36)

Note that $b(t)$, the fluid limit for the number of blocked customers in Station 1, is given by
\begin{align}
b(t) &= \sum_{i=2}^{k} (x_i(t) - N_i)^+, \quad t \geq 0.
\end{align}
(37)

**Appendix C: Proof of Theorem 1**

The function $C(N)$ in (14) equals
\begin{align}
C(N) &= constant - (C_o + C_u) \int_{0}^{N} [f(x) - Z] dx,
\end{align}
(38)
where
\begin{align}
f(x) &= \int_{0}^{T} 1_{\{r_d(t) \geq x\}} dt \quad \text{and} \quad Z = \frac{C_o T}{C_o + C_u}.
\end{align}
(39)

Therefore, it suffices to prove that the function $F(N)$, given by
\begin{align}
F(N) &= \int_{0}^{N} [f(x) - Z] dx,
\end{align}
(40)
is maximized by $N^*$ in (16).

Note that $f(x)$ is non-negative and non-increasing in $x$, where $f(0) = T$ and $\lim_{x \to \infty} f(x) = 0$. In addition, $Z \in [0, T]$, hence $f(x)$ crosses level $Z$. The function $F(N)$, for $N$ starting from 0, is first an integral of a non-negative integrand, hence is increasing in $N$. Then, after the first $N$ for which $f(N) = Z$, it is decreasing.

This proves that $F(N)$ is maximized (globally) at point $N$, where $f(N) = Z$.

We conclude the proof by showing that $N^*$ in (16) satisfies $f(N^*) = Z$. Substituting $N^*$ into (39) gives
\begin{align}
f(N^*) &= \int_{0}^{T} 1_{\{r_d(t) \geq r_d(Z)\}} dt = \int_{0}^{T} 1_{\{t \leq Z\}}(t) dt = Z,
\end{align}
(41)
since $r_d$ is a decreasing function. Therefore, $N^* = r_d(Z)$, as in (16).

**Remark 3.** When $r_d$ is continuous and strictly decreasing, $f(x)$ is in fact its inverse $r_d^{-1}$.

\qed

**Appendix D: Choosing the Candidate Solution**

We now describe the method that motivates $N^*$, as in (16), to be a natural candidate for maximizing $C(N)$ in (14). This method requires additional assumptions about $r(t)$, $r_d(t)$ and $\lambda$. Theorem 1, though, does not make these assumptions and is, therefore, more general.

Figure 8 shows an illustration of the overage and underage periods for a specific number of beds ($N = 280$): on the left, according to $r(t)$ and on the right according to $r_d(t)$. The light areas mark underage periods, where the offered load is higher than the number of beds. The dark areas mark overage periods. The areas of each color are equal in the two figures.
We assume that \( r_d(t) \) is an invertible function and define \( t^* \) to be the intersection point between \( r_d(t) \) and \( N \) such that \( r_d(t^*) = N \); then \( t^* = r_d^{-1}(N) \). We can rewrite \( C(N) \) to get

\[
C(N) = C_u \int_0^{r_d^{-1}(N)} [r_d(t) - N] dt + C_o \int_{r_d^{-1}(N)}^T [N - r_d(t)] dt.
\]

(42)

Now assume that \( r_d^{-1}(N) \) is a continuous differential function and differentiate Equation (42) according to Leibniz’s differentiation rule:

\[
\dot{C}(N) = C_o (T - r_d^{-1}(N)) - C_u r_d^{-1}(N) = -(C_o + C_u) r_d^{-1}(N) + C_o T.
\]

(43)

Since \( C(N) \) approaches \( \infty \) as \( N \) approaches \( \infty \) and achieves a high positive value for \( N = 0 \), we minimize \( C(N) \) by equating the derivative to 0. This gives rise to

\[
r_d^{-1}(N) = \frac{C_o T}{C_o + C_u}.
\]

(44)

Applying \( r_d \) to both sides yields the optimal \( N^* \) in Equation (16).

Since \( C_o \) and \( C_u \) are non-negative numbers and \( r_d^{-1}(N) \) is decreasing in \( N \), \( \dot{C}(N) \) is monotonically non-decreasing, and therefore, \( C(N) \) is convex and \( N^* \) in Equation (16) minimizes \( C(N) \).

**Appendix E: Proof of Theorem 2**

In our proof, we use the following proposition, which is proved in Appendix F:

**Proposition 2.** \( C(N) \) in (14) is a convex function.

We solve problem (23) for the case where \( N \leq B \), and for the case where \( N \geq B \). Then, we choose the solution which minimizes the overall cost. The option for \( N = B \) is included in both cases since their solutions are identical.

**Step 1:** Find \( N_k^1 \), the optimal number of beds if no new beds are added, by solving \( C_K(N) \) for \( N \leq B \).
Since $C(N)$ is a convex function, if the optimal solution for the unconstrained problem is in the allowed range (i.e., $N^* \leq B$), then this will be the solution for the constraint problem as well. If not, the solution will be at the edge of the range. Formally:

$$N^1_k = \begin{cases} r_d \left( \frac{C_o T}{C_o + C_u} \right), & r_d \left( \frac{C_o T}{C_o + C_u} \right) \leq B \\ B, & \text{otherwise.} \end{cases}$$  \hspace{1cm} (45)

**Step 2:** Find $N^2_k$, the optimal number of beds, where $(N - B)$ new beds are added, by solving $C_K(N)$ for $N \geq B$, as follows:

$$\begin{align*}
\text{minimize} & \quad C(N) + K(N - B) \\
\text{subject to} & \quad -N + B \leq 0.
\end{align*}$$  \hspace{1cm} (46)

Since the objective function remains convex, we solve the unconstrained problem and check whether the solution is in the allowed range. For this, we use the following statement:

The optimal solution, which minimizes the unconstrained problem

$$C^{(u)}_K(N) = C(N) + K(N - B),$$  \hspace{1cm} (47)

is given by

$$N^{(u)*}_K = r_d \left( \frac{C_o T + K}{C_o + C_u} \right).$$  \hspace{1cm} (48)

This is because the function $C^{(u)}_K(N)$ in (47) can be written in the same structure as in (38) for

$$C = \frac{C_o T + K}{C_o + C_u}.$$  \hspace{1cm} (49)

In order to justify the introduction of new beds, we must have $K \leq TC_u$, and therefore, $0 \leq C \leq T$. Since $0 \leq f(x) \leq T$, $f(x)$ crosses $C$ and the proof in Theorem 1 holds. The optimal solution for (47) is $N^{(u)*}_K = r_d(C)$, as in (48).

The solution for (46) is, therefore,

$$N^2_k = \begin{cases} r_d \left( \frac{C_o T + K}{C_o + C_u} \right), & r_d \left( \frac{C_o T + K}{C_o + C_u} \right) \geq B \\ B, & \text{otherwise.} \end{cases}$$  \hspace{1cm} (50)

**Step 3:** Combining the results of Steps 1 and 2, yields the solution in Equation (24).

**Appendix F: Proof of Proposition 2**

It is sufficient to prove that $F(N)$ in (40) is a concave function. According to Sierpinski’s Theorem (see Donoghue 1969), a midpoint concave function that is continuous is, in fact, concave. Since the function $F(N)$ is an integral of $N$, and therefore, continuous, it is sufficient to prove that it is midpoint concave. Without loss of generality, it suffices to prove midpoint concavity by proving that for every $N \geq 0$,

$$F(N/2) \geq \frac{F(N)}{2}.$$  \hspace{1cm} (51)

In other words, we need to prove that
\[ 2 \int_0^{N/2} [f(x) - C] \, dx \geq \int_0^N [f(x) - C] \, dx, \quad (52) \]

which is equivalent to proving that

\[ 2 \int_0^{N/2} f(x) \, dx \geq \int_0^N f(x) \, dx. \quad (53) \]

Since \( f \) is a non-increasing non-negative function, we must have

\[ 2 \int_0^{N/2} f(x) \, dx \geq \int_0^{N/2} f(x) \, dx + \int_{N/2}^N f(x) \, dx = \int_0^N f(x) \, dx, \quad (54) \]

which completes the proof. \( \square \)