Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Models and Insights for Hospital Inpatient Operations: Time-Dependent ED Boarding Time

Pengyi Shi, Mabel C. Chou, J. G. Dai, Ding Ding, Joe Sim

To cite this article:
Pengyi Shi, Mabel C. Chou, J. G. Dai, Ding Ding, Joe Sim (2016) Models and Insights for Hospital Inpatient Operations: Time-Dependent ED Boarding Time. Management Science 62(1):1-28. http://dx.doi.org/10.1287/mnsc.2014.2112

MANAGEMENT SCIENCE
Vol. 62, No. 1, January 2016, pp. 1–28
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2014.2112

Models and Insights for Hospital Inpatient Operations: Time-Dependent ED Boarding Time

Pengyi Shi
Krannert School of Management, Purdue University, West Lafayette, Indiana 47907, shi178@purdue.edu

Mabel C. Chou
Department of Decision Sciences, NUS Business School, National University of Singapore, Singapore 119245, mabelchou@nus.edu.sg

J. G. Dai
School of Operations Research and Information Engineering, Cornell University, Ithaca, New York 14853, jim.dai@cornell.edu

Ding Ding
School of International Trade and Economics, University of International Business and Economics, Beijing 100029, China, dingd.cn@gmail.com

Joe Sim
NUS Yong Loo Lin School of Medicine and NUS Business School, National University of Singapore, Singapore 119245; and National University Hospital, Singapore 119074, joe_sim@nuhs.edu.sg

One key factor contributing to emergency department (ED) overcrowding is prolonged waiting time for admission to inpatient wards, also known as ED boarding time. To gain insights into reducing this waiting time, we study operations in the inpatient wards and their interface with the ED. We focus on understanding the effect of inpatient discharge policies and other operational policies on the time-of-day waiting time performance, such as the fraction of patients waiting longer than six hours in the ED before being admitted. Based on an empirical study at a Singaporean hospital, we propose a novel stochastic processing network with the following characteristics to model inpatient operations: (1) A patient's service time in the inpatient wards depends on that patient's admission and discharge times and length of stay. The service times capture a two-time-scale phenomenon and are not independent and identically distributed. (2) Pre- and post-allocation delays model the extra amount of waiting caused by secondary bottlenecks other than bed unavailability, such as nurse shortage. (3) Patients waiting for a bed can overflow to a nonprimary ward when the waiting time reaches a threshold, where the threshold is time dependent. We show, via simulation studies, that our model is able to capture the inpatient flow dynamics at hourly resolution and can evaluate the impact of operational policies on both the daily and time-of-day waiting time performance. In particular, our model predicts that implementing a hypothetical policy can eliminate excessive waiting for those patients who request beds in mornings. This policy incorporates the following components: a discharge distribution with the first discharge peak between 8 A.M. and 9 A.M. and 26% of patients discharging before noon, and constant-mean allocation delays throughout the day. The insights gained from our model can help hospital managers to choose among different policies to implement depending on the choice of objective, such as to reduce the peak waiting in the morning or to reduce daily waiting time statistics.

Keywords: inpatient flow management; early discharge; time-dependent waiting time; stochastic network model; ED boarding
History: Received August 26, 2012; accepted April 15, 2014, by Assaf Zeevi, stochastic models and simulation. Published online in *Articles in Advance* April 22, 2015.

1. Introduction

Inpatient beds are one of the most critical resources in hospitals. Inpatient flow and bed management has crucial impacts on hospital operations (Hall 2012), especially on emergency department (ED) crowdedness (Litvak et al. 2001, Howell et al. 2008, Bair et al. 2010, Wong et al. 2010, Schneider et al. 2001). Prolonged waiting time for admission to inpatient wards, also known as ED boarding, has been identified as a key contributor to ED overcrowding worldwide (United States General Accounting Office 2003, Hoot and Aronsky 2008, Pines et al. 2011b). This paper aims to provide a high-fidelity model to capture the dynamics of inpatient flow with a particular focus on

predicting the time-of-day waiting time performance during the process of transferring from the ED to wards and identifying strategies (from the inpatient side) to reduce the waiting. Although the model is built on an extensive empirical study at one Singaporean hospital, we believe that the modeling framework can be adapted to other hospitals, based on the similarity in many empirical observations between this hospital and others.

1.1. Motivation and Research Questions

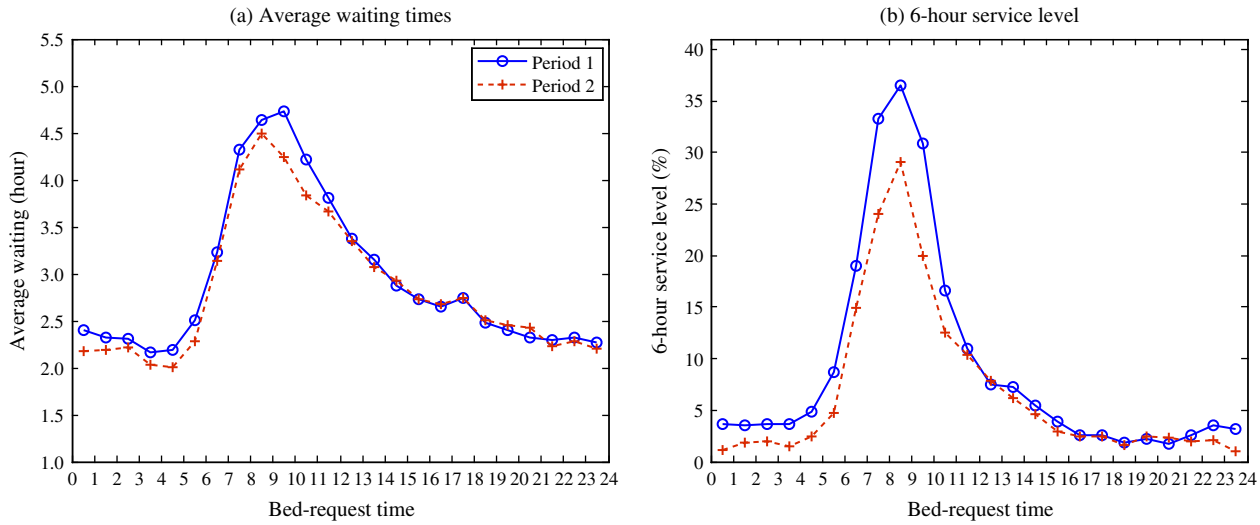
National University Hospital (NUH) is one of the major public hospitals in Singapore. It operates a busy ED and a large inpatient department that has approximately 1,000 beds to serve patients admitted from the ED and other sources. At NUH, approximately 20% of patients visiting the ED are admitted into a general ward (GW) after finishing the treatment in ED, thereby becoming ED-GW patients. The waiting time for admission to a ward of an ED-GW patient, (simply the waiting time hereafter) is defined as the duration between the time when ED doctors made the decision to admit the patient (i.e., the bed-request time of the patient) and the time when the patient was admitted to a GW.

1.1.1. Time-of-Day Waiting Time Performance. From January 1, 2008, to June 30, 2009, called period 1 in this paper, the average waiting time at NUH was 2.82 hours (169 minutes), which does not seem to be very long. However, this level of complacency immediately evaporates if we examine the waiting times of patients requesting beds in mornings. The solid curve in Figure 1(a) shows that the average waiting time is more than 4 hours long for patients who request a bed between 7 A.M. and 10 A.M. Moreover, among these patients, more than 30% of them have to wait 6 hours or longer, as shown in Figure 1(b). In this paper, we define the 6-hour service level as the fraction of patients who have to wait 6 hours or longer.

Although no patient likes any wait, 6 hours or more is extremely undesirable, not only because patients can get very frustrated during the long wait (Pines et al. 2008), but also because of the adverse outcome associated with it. Liu et al. (2009) and Singer et al. (2011) have discovered that patients who waited longer than 6 hours after their admission decisions have been made are more likely to experience longer inpatient stays, higher mortality rates, and other undesirable events in the ED such as suboptimal blood pressure control. In addition, patients continue to occupy ED resources while waiting to be transferred to wards and can block new patients from being treated in the ED, which can lead to ED overcrowding and sometimes ambulance diversion (Allon et al. 2013). Thus, it is important for hospitals to eliminate the excessive amount of waiting, especially for morning bed requests.

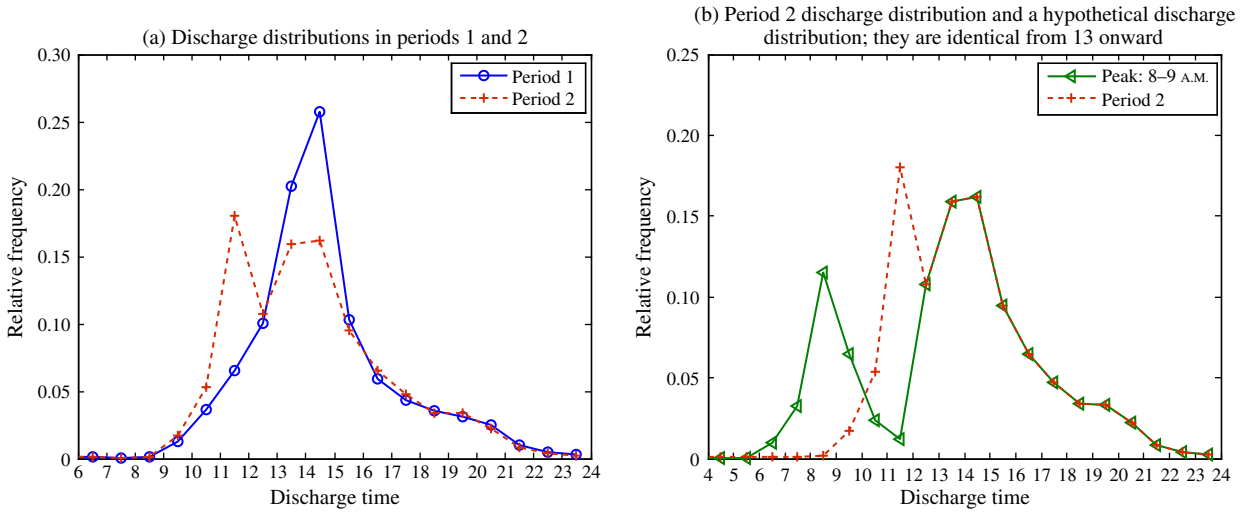
1.1.2. Discharge Pattern and Early Discharge Policy. The inpatient discharge policy is believed by NUH to have contributed to the prolonged waiting times for ED-GW patients requesting beds in the morning. The solid curve in Figure 2(a) plots the discharge distribution of patients from GWs at NUH in period 1. Clearly, the peak discharge hour is between 2 P.M. and 3 P.M. Therefore, many admissions must wait until after 3 P.M., although bed requests of ED-GW patients can occur during the entire day (see the solid curve in Figure 7 in §4.1). In other words,

Figure 1 **(Color online) Hourly Waiting Time Statistics for ED-GW Patients**



Notes. Period 1: January 1, 2008, to June 30, 2009; period 2: January 1, 2010, to December 31, 2010. Each dot represents the average waiting time or 6-hour service level for patients requesting beds in that hour. For example, the dot between 7 and 8 represents the value of the hourly statistics between 7 A.M. and 8 A.M. The 95% confidence intervals are reported in Table 1 of the online supplement.

Figure 2 (Color online) Discharge Time Distributions in Periods 1 and 2 and a Hypothetical Discharge Distribution



Notes. In the hypothetical discharge distribution, the first peak is between 8 A.M. and 9 A.M., and 26% of patients are discharged before noon. Each dot represents the fraction of patients who are discharged during that hour. The values in the first four hours are nearly zero in all three distributions and are not displayed.

if there is no bed immediately available for a morning bed request, the incoming patient is likely to wait until afternoon to be admitted.

In fact, the time dependency of waiting times is not unique to NUH. Similar waiting time curves have been observed in other hospitals (see Figure 16 of Armony et al. 2015), and so have the number of patients waiting at different times of the day (Powell et al. 2011, Hall 2012). Meanwhile, the bed-request and discharge patterns in many other hospitals are also similar to what we observed at NUH; see, e.g., Figure 1 of Powell et al. (2011) and Figure 10 of Armony et al. (2015). Studies in the literature (Birjandi and Bragg 2008, Yancer et al. 2006) and by government agencies (e.g., Department of Health, United Kingdom 2004) have recommended discharging patients at earlier hours of the day to eliminate the temporary mismatch between bed demand and supply in the morning.

In July 2009, NUH itself launched an early discharge campaign. After six months' implementation, a new discharge pattern emerged in period 2: January 1, 2010, to December 31, 2010. The dashed curve in Figure 2(a) displays the new discharge distribution. A morning discharge peak arises, occurring between 11 A.M. and noon; 26% of the patients are discharged before noon in period 2, doubling the proportion in period 1 (13%). The daily average waiting time is reduced from 2.82 hours (169 minutes) in period 1 to 2.77 hours (166 minutes) in period 2, and the daily 6-hour service level is reduced from 6.52% in period 1 to 5.13% in period 2. The dashed curves in Figures 1(a) and 1(b) plot the time-dependent hourly average waiting time and 6-hour service level in period 2, respectively. From these empirical results, we observe that (a) some improvement in reducing the peak hourly 6-hour service level has been achieved in period 2, and (b) little progress has been made in eliminating the long waiting times for morning bed requests (flattening the hourly waiting time statistics) or reducing the daily waiting time statistics.

These empirical observations raise two issues. First, it is unclear whether the improvements in period 2 result from the NUH's early discharge campaign. As in many hospitals, the operating environment is continuously changing at NUH. Bed capacity is being increased in response to the rising number of patients seeking treatment. In period 2, the bed occupancy rate (BOR) has reduced by 2.7% (Shi et al. 2014). Therefore, it is difficult to evaluate the impact of the early discharge policy through empirical analysis alone. Second, one wonders if there is any discharge policy, perhaps combined with other operational policies, that can achieve a more significant improvement in flattening or reducing the waiting time statistics. Unfortunately, it is prohibitively expensive for hospitals to experiment with various options in a real operational environment to identify such policies. Therefore, we need a high-fidelity model to (i) capture the inpatient flow dynamics and predict the time-dependent waiting time performance and (ii) quantify the impact of operational policies such as early discharge and identify strategies to eliminate the long waiting times.

1.2. Contributions

This paper makes two major contributions to the modeling and practice of inpatient flow management.

1.2.1. Modeling. For the first contribution, we develop a new stochastic network model that reproduces, at high fidelity, many empirical performance measures at both the hospital and the medical specialty levels. In particular, the model can approximately replicate the time-dependent hourly waiting time performance. For the model to be able to capture the inpatient operations at hourly resolution, we find that several key features must be built in. They include a two-time-scale service time model, an overflow mechanism among multiple server pools, and pre- and post-allocation delays that capture the extra amount of delay caused by resource constraints other than bed unavailability during the ED-to-wards transfer process. Under our two-time-scale service time model, service times of inpatients are not independent and identically distributed (iid). We will elaborate this service time model and other key features in §3. Time-varying $M_t/GI/n$ queues or their network versions, where the arrival process is Poisson with a time-varying arrival rate and the service times are iid, have been used in the literature to model hospital operations; see, for example, Green (2002), Koizumi et al. (2005), Allon et al. (2013). Despite our best efforts, we are not able to reproduce the time-dependent performance curves using these models. See §5.2 for simulation results for models that miss each one of the three key features.

Our model strikes a proper balance between analytical tractability and fidelity, although we have mainly used simulation to generate insights in this paper. Indeed, in a preliminary work (Dai and Shi 2014), the authors are able to analyze some simplified versions of the proposed model while still keeping certain key features, including the two-time-scale service time model and allocation delays.

We want to emphasize that studying inpatient flow dynamics at hourly resolution and capturing time-of-day performance are important, especially when one evaluates policies that impact the interface between the ED and wards, where hours of waiting matter. For example, our model predicts that certain types of discharge policies can significantly reduce waiting times for morning bed requests but have limited impact on the daily waiting time statistics (see also the second contribution below). By studying the time-of-day performance, we are able to gain insights into the impact of such policies on certain subgroups of patients, in addition to the aggregated impact on all patients. Moreover, as pointed out by Armony et al. (2015), understanding the system's behavior at hourly resolution is of particular importance for operational planning when nurses and physicians are modeled as servers, e.g., for planning nurse staffing. Thus, our model can potentially be used to aid other operational decisions that require an understanding of the time-varying dynamics of inpatient flow.

1.2.2. Practice. through simulation analysis of the proposed model, we obtain managerial insights into the impact of early discharge and other operational policies on both the daily and time-of-day waiting time performance. First, consistent with the empirical observations, the period 2 early discharge alone has little impact on reducing or flattening the waiting time of ED-GW patients. Second, if the hospital is able to (i) move the first discharge peak in period 2 three hours earlier to occur between 8 A.M. and 9 A.M. (see the solid curve in Figure 2(b)) and (ii) meanwhile stabilize the time-varying allocation delays, then the hourly waiting time curves can be approximately flattened (see Figure 17 in §6.2). However, the daily waiting time statistics still show limited reductions. Third, we identify policies that can significantly impact the daily waiting time performance such as increasing bed capacity or reducing the mean allocation delays; these policies do not necessarily flatten the hourly waiting time curves, though.

The different impacts on the daily and hourly waiting time performance of these policies can be explained by the phenomenon of separation of time scales, which is captured through our new service time model. A patient's waiting time essentially comes from two sources: one is the mismatch between the daily number of arrivals and discharges, and the other is the mismatch between the discharge timing and the hourly arrival pattern. Early discharge policy can reduce the second type of wait, but not the first; thus it mainly affects the hourly waiting time performance. In contrast, increasing capacity mainly reduces the first type of wait and, thus, affects the daily performance. These insights can provide useful guidelines for hospital managers on choosing among different policies based on their objectives. See detailed discussion in §6.5.

To the best of our knowledge, this paper is the first to build a stochastic model to analyze the effect of discharge policy in combination with other strategies such as stabilizing allocation delays. The most relevant paper is a recent one by Powell et al. (2011), where the authors propose a deterministic fluid model to analyze the effect of discharge timing on the waiting time for admission to wards. Their model provides a simple method to calculate the hourly mean patient count (number of patients in service and waiting), and this method can actually be supported by a more rigorous study in an ongoing work (Dai and Shi 2014) based on the two-time-scale service time model proposed in this paper. However, the fluid method is not enough to calculate the mean queue length or other performance measures that depend on the entire distribution of the hourly patient count. Therefore, some of the managerial insights generated in Powell et al. (2011) can be misleading. For example, the authors

find that by shifting the peak inpatient discharge time four hours earlier, the waiting time can be reduced to zero; but zero waiting can hardly be achieved in any hospital with as much as 90% bed utilization and random arrivals and service times. We believe our model is more comprehensive and sophisticated so that it captures inpatient flow operations at hourly resolution and generates insights on many operational policies including discharge timing. Some other relevant works on discharge policies are mostly empirical studies. For example, Khanna et al. (2011) classify admission data from 23 Australian hospitals into five categories based on the relative timing of daily admission and discharge curves and use statistical analysis to show that days with late discharge peaks contribute significantly to ED overcrowding.

1.3. Literature Review and Paper Outline

Hospital patient flow has been studied extensively in the operations research literature. For example, Armony et al. (2015) and Hall et al. (2006) conduct detailed studies of patient flow in various departments at an Israeli and a U.S. hospital, respectively. Readers are also referred to the many articles cited in these two papers for further references. Armony et al. (2015) do not focus on discharge policies, but they empirically study the transfer process flow from ED to GW (which they call "internal wards"). Discrete-event simulation and queueing theory are two commonly used approaches for modeling and improving patient flow (Green 2006, Jacobson et al. 2006, Zeltyn et al. 2011). Compared to the rich literature on patient flow models of EDs, inpatient flow management and the interface between the ED and inpatient wards have received less attention; see the same discussion in Section 4 of Armony et al. (2015). Related works on inpatient operations include capacity allocation and flow improvement in specialized hospitals or wards (Griffin et al. 2011, Cochran and Bharti 2006, de Bruin et al. 2007, Green 2002), ward nurse staffing (Vericourt and Jennings 2011, Yankovic and Green 2011), bed assignment and overflow (Thompson et al. 2009, Mandelbaum et al. 2012), and elective admission control and design (Helm and Van Oyen 2014, Helm et al. 2011). Note that Yankovic and Green (2011) demonstrate that the admission or discharge blocking caused by nurse shortages can have a significant impact on system performance. This insight is consistent with our findings on allocation delays.

Stochastic network models have been a common tool to study manufacturing, communications, and service systems (Gans et al. 2003, Bertsekas and Gallager 1992, Yao 1994). In particular, research motivated by call center operations has extensively studied stochastic systems with time-varying arrivals and time-dependent performance. For example, Feldman et al. (2008) and Liu and Whitt (2012) propose staffing algorithms to achieve time-stable performance. Unlike call center models, our hospital model has extremely long service times with an average of approximately five days. Within the service time of a typical patient, the arrival pattern has gone through five cycles. Therefore, existing approximation methods developed for call center models are not applicable to our hospital model. Moreover, the servers in our model are inpatient beds. It is not realistic to adjust the number of beds within a short time window.

The remainder of this paper is organized as follows. In §2, we give a brief description of the NUH inpatient department and the performance measures on which we focus. In §3, we introduce the general framework of our proposed stochastic network model, which captures the inpatient flow operations. In §4, we populate the proposed stochastic network model with NUH data. In §5, we verify the populated model by comparing the model output with empirical performance. In §6, we use the populated model to generate a number of managerial insights for reducing and flattening waiting times for admission to wards. The paper concludes in §7.

2. NUH Inpatient Department

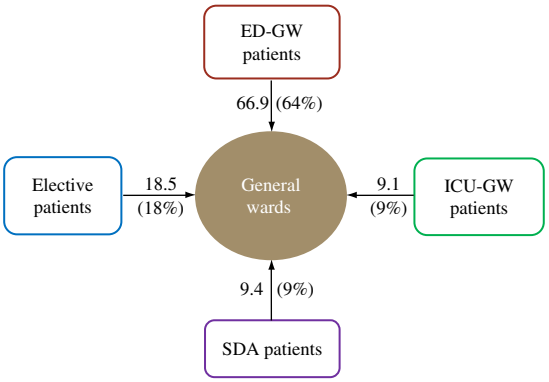
This section briefly describes the operations of the NUH inpatient department. We focus on 19 GWs, excluding a certain number of wards including intensive-care-unit (ICU) wards, isolation wards, high-dependence wards, pediatric wards, and obstetrics and gynecology (OG) wards. A bed in a GW is called a general bed or sometimes a "floor bed" in U.S. hospitals. The total number of general beds at NUH ranged from 555 to 638 between January 1, 2008, and December 31, 2010. The precise definition of GW and reasons we exclude other wards from GWs are presented in the companion paper (Shi et al. 2014).

2.1. Admission Sources

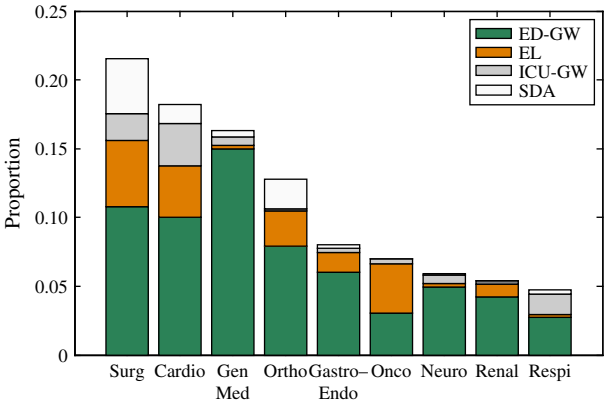
Patients admitted to the general wards are mainly from four sources. They are ED-GW, ICU-GW, elective (EL), and same-day-admission (SDA) patients. ED-GW patients are those who have completed treatment in the ED and need to be admitted to a GW. ICU-GW patients are those patients who are initially admitted to ICU-type wards (from either the ED or other external sources) and are later transferred to GWs. Most of the EL and SDA patients come to the hospital to receive elective surgeries, and they usually have less urgent medical conditions than ED-GW or ICU-GW patients. The difference between EL and SDA patients is that EL patients are usually admitted to a GW in the afternoons before the day of surgery, whereas SDA patients first go to the operating room to receive surgery (usually in the morning). After the

Figure 3 **(Color online) Four Admission Sources to GWs and Nine Medical Specialties**

(a) Admission sources and daily admission rates



(b) Patient distribution based on medical diagnosis



Note. Daily admission rates and patient distributions are estimated from data from periods 1 and 2.

surgery, SDA patients stay temporarily in the SDA ward, typically for a few hours, and then are admitted to a GW. Therefore, it is expected that an EL patient typically stays in a GW bed at least one day longer than a SDA patient.

Figure 3(a) shows the four admission sources and their average daily admission rates, which are estimated from combining the data from periods 1 and 2. Each patient is only counted once when we calculate the admission rate for the corresponding admission source, although some patients may be transferred out of and back into GWs after the initial admission. In this paper, patients admitted to GWs from any of the four sources are called "general patients."

2.2. Medical Specialties

General patients are classified by one of nine medical specialties based on diagnosis at the time of admission as an inpatient: surgery, cardiology, general medicine, orthopedics, gastroenterology and endocrinology, oncology, neurology, renal disease, and respiratory disease. Although gastroenterology and endocrinology are two different medical specialties, in this paper we group them together and denote them as Gastro-Endo. The grouping is based on the fact that patients from these two specialties share the same ward and have similar length-of-stay (LOS) distributions. See Teow et al. (2012) for the same classification. We group dental, eye, and ENT patients into surgery for similar reasons. As we mentioned at the beginning of §2, two other specialties, OG and pediatrics, are excluded from our study.

Figure 3(b) plots the distribution of general patients among different specialties and admission sources. There is no significant difference in the patient distribution between two periods, so we plot the figure using the combined data. Different specialties show very different admission-source distributions. For example, the majority of general medicine patients are admitted from the ED, whereas a significant proportion of surgery patients are EL and SDA patients.

2.3. Performance Measures

2.3.1. Waiting Time. Recall that we define the waiting time of an ED-GW patient as the duration between the bed-request time and actual admission time. In §1, we empirically compare the daily and hourly waiting time statistics in period 1 with those in period 2. Our definition of waiting time is consistent with the convention in the medical literature (United States General Accounting Office 2003, Sinagpore Ministry of Health 2012), except that we use the admission time to wards as the end point of the waiting period whereas the literature usually uses the time when the patient exits the ED. Thus, our reported waiting time is a slight overestimation of the value computed in the conventional way. (The gap between a patient's exiting the ED and admission to a ward is approximately 18 minutes on average at NUH.)

For an ICU-GW or an SDA patient, although there is a delay between the bed-request time and the departure time from the original ward, this waiting time is taken less seriously than that of ED-GW patients at NUH. This claim is supported by our empirical observations that the average waiting time is more than 7 hours for ICU-GW patients and approximately 3.5 hours for SDA patients, both longer than that of ED-GW patients (with an average of less than 3 hours). The major reason could be that the ICU-GW and SDA patients have been receiving care at their current wards, so this waiting time is not an issue unless there is a bed shortage in the ICU-type wards or the SDA ward. In this paper, we focus on the waiting time for ED-GW patients.

The waiting time statistics for ED-GW patients for different medical specialties are different. Generally speaking, renal patients show the longest average waiting time, and their 6-hour service level is more than 10%. Surgery, general medicine, and respiratory patients have better performance on the waiting time statistics than other specialties. Table 5 in §5.1 displays the average waiting time and 6-hour service level of each specialty in period 1.

2.3.2. Overflow Proportion and Other Performance. In NUH, each general ward is designated to serve patients from one or more specialties. Usually patients are admitted to the designated wards, which we call the "primary wards." However, when an ED-GW patient has waited for several hours in the ED, but no bed from the primary wards is available or expected to be available in the next few hours, NUH may "overflow" the patient to a nonprimary ward as a temporary expedient. Such overflow events may also occur among patients admitted from other sources; for example when ICU-type wards need to free up capacity, ICU-GW patients may be overflowed. In this paper, we define the overflow proportion as the number of patients admitted to nonprimary wards divided by the total number of admissions.

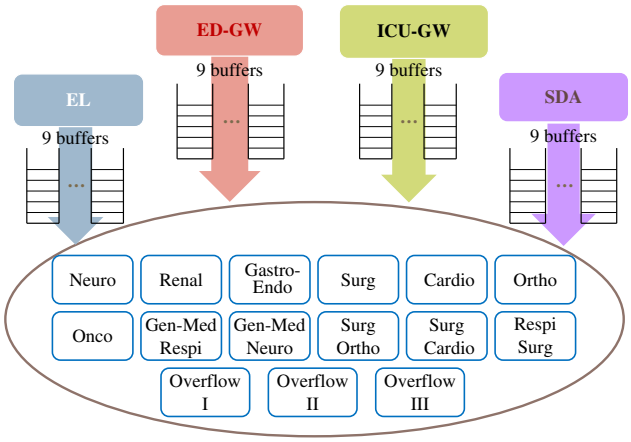
Obviously, there is a trade-off between patient waiting time and overflow proportion. On the one hand, the waiting time can always be reduced by overflowing patients more aggressively, since overflow acts as resource pooling. On the other hand, overflow decreases the quality of care delivered to patients and increases hospital operational costs (Teow et al. 2012). In NUH, the average overflow proportion among all patients is 26.95% and 24.99% for periods 1 and 2, respectively. The overflow proportion for all ED-GW patients is 29.91% in Period 1 and 28.54% in period 2, slightly higher than the values for all patients. The lower overflow proportion in period 2 indicates that the reduced waiting time for ED-GW patients in period 2 does not result from a more aggressive overflow policy. Readers are referred to §2.3 of Shi (2013) for discussion of specialty-level and ward-level overflow proportions.

Besides the waiting time and overflow proportion, other performance measures of interest to us include (a) the queue length, which counts the number of ED-GW patients waiting in the ED; and (b) bed utilization, which is the proportion of beds being occupied by patients over all beds.

3. A Stochastic Network Model for Inpatient Operations

In this section, we describe a general framework of our proposed stochastic model, which is built upon an extensive empirical study of NUH inpatient operations (Shi et al. 2014) but which could be adapted to other hospitals. We first overview the basic ingredients of the stochastic processing network and the basic patient flow in §3.1. Then in §§3.2–3.4, we specify the details of three modeling features that are critical to capture inpatient operations. These features are a non-iid, two-time-scale service time model, an overflow mechanism, and pre- and post-allocation delays that create additional delay during a patient's admission process. Finally, we discuss service policies in §3.5.

Figure 4 (Color online) Arrival and Server Pool Configuration in the Stochastic Model of NUH Inpatient Department



Under a specified service policy and a specification of input parameters estimated from a hospital data set, the proposed stochastic model can be populated and simulated on a computer. Section 4 details how we populate the model using NUH data. Section 5 verifies the populated model by comparing the simulation output against the empirical estimates. We will see that our proposed stochastic model can approximately replicate waiting time performance, even at hourly resolution, from the empirical data.

3.1. A Stochastic Processing Network with Multiserver Pools

A stochastic processing network processes incoming customers (patients) of various classes. The basic ingredients of a stochastic processing network are servers, buffers, activities, and service policies. A general stochastic processing network was proposed in Harrison (2003) and precisely specified in Dai and Lin (2005). Our proposed stochastic model is a special stochastic processing network that has a single-pass routing structure: after a service completion of a customer, the customer departs from the system permanently. Figure 4 depicts a stochastic processing network representation of the NUH inpatient department.

3.1.1. Servers. In this paper, general ward beds play the role of servers, and these servers are grouped into J parallel server pools. Each server pool models a ward or a group of similar wards. We use n_j to denote the number of servers in pool j, $j = 1, \ldots, J$. These n_j servers are assumed to be identical.

3.1.2. Customers. The J server pools serve customers from K different classes. The customers in the model correspond to patients who need to receive hospital care in a general ward, and a customer class can be a combination of an admission source and a medical specialty, sometimes with other criteria such as admission time. Customers in the same class are homogeneous in the sense that they follow the same arrival process, service time specification, and service priority. In our model, each customer admission source is associated with an arrival process, which is used to model the patient bed-request process. Each arriving patient (from any of the admission sources) is then assigned to a medical specialty with a certain probability that depends both on the admission source and on the arrival hour. Hereafter, we use patient and customer, bed request and arrival, and bed and server interchangeably.

3.1.3. Buffers. Upon arrival, each patient is first held in a buffer, waiting to be assigned a bed and later to be admitted into the bed. The buffer can be a bed in the ED or another area for a patient to wait before admission to a GW. The patients waiting in these buffers are processed following certain priorities that are specified by a service policy as described below.

3.1.4. Activities and service policies. Each server pool is designated to serve patients from one or more medical specialties, and we call the pool a "primary pool'' for patients from the designated specialties. We assume that each class of patients can potentially be assigned to any of the J server pools in the model. If a patient is assigned to a primary server pool, we say that patient is "right-sited,'' otherwise, that patient is "overflowed." Adapting the stochastic processing network terminology to the hospital setting, an activity is the binding of a server pool serving a particular class of patients. When the server pool is a primary pool for the class, the corresponding activity is said to be a primary activity. Clearly, primary activities are more desirable because they avoid patient overflow. However, to reduce waiting time, it is sometimes necessary to activate nonprimary activities. A service policy dictates which activities should be initiated at a decision time point. In the hospital setting, a service policy is also known as a bed assignment policy that dictates which beds should be assigned to which waiting patients at a decision time point. The decision time points have three categories: the arrival time of a patient, the departure time of a patient, and the overflow trigger time of a patient. A patient can be overflowed only when the waiting time exceeds the preassigned overflow trigger time. The service policy also dictates the choice of the overflow trigger time for each patient.

3.1.5. Basic patient flow. After a bed is assigned to a patient, that patient has to experience extra delays (pre- and post-allocation delays) before being admitted to the bed. Thus, a patient's admission time is different from the bed assignment time in our model. Once a patient is admitted, that patient occupies the bed until departure. The duration of occupation is called the patient's "service time.'' The service time of each patient is random and follows the two-time-scale model (1) below. At the end of the service time, the patient departs from the system. Thus, our proposed stochastic network model has a single-pass structure. The departure times for most patients in our model correspond to their discharge times from the hospital, and hereafter we use departure and discharge interchangeably.

3.2. Critical Feature 1: A Two-Time-Scale Service Time Model

The service time, S, of a patient is the duration between the admission time and the discharge time. We use day as the time unit for service times unless specified otherwise. Clearly, the service times of patients are random. Both the patient's medical condition and hospital operational policies can affect the service time. We adopt the following model to separate different sources of influence on service times:

$$S = LOS + h_{\text{dis}} - h_{\text{adm}}. \tag{1}$$

We will discuss the rationale for using service time model (1) in §3.2.2 below. Here, LOS stands for "length of stay'' and is equal to the number of midnights that the patient spends in a ward or, equivalently, day of discharge minus day of admission, and h_{dis} and h_{adm} stand for the time of day when the patient is admitted and discharged, respectively. The time of day is between 0 and 1, with midnight being 0 day and 12 P.M. (noon) being 0.5 day. For a patient who is discharged on the same day of admission, our definition of that patient's LOS is equal to 0, whereas when hospitals report occupancy level or some other statistics (Hall et al. 2010, Centers for Disease Control and Prevention 2011), the LOS of such same-day discharge patients is adjusted to 1 for accounting and cost recovery purposes.

3.2.1. Non-iid Service Times. Based on an extensive empirical study (Shi et al. 2014), we make the

following assumptions for the service time model in (1):

(a) The discharge hour h_{dis} is independent of LOS and of h_{adm}; § 8.5 of Shi et al. (2014) provides some empirical evidence for this assumption.

(b) LOS distributions are class dependent. Patients from different medical specialties or admission sources follow different LOS distributions.

(c) For each class of patients, their LOS forms a sequence of iid random variables following a discrete distribution. One can use an empirical LOS distribution directly estimated from data or a discrete version of the log-normal distribution based on our empirical fitting results (see Figure 8 in §4.4) and similar findings in Armony et al. (2015).

(d) The discharge hours h_{dis} for each class of patients form another sequence of iid random variables following a certain discharge distribution. See Figure 2(a) for an example of NUH's discharge distribution.

(e) We assume all iid sequences of LOS and h_{dis} are independent of each other; i.e., there is no dependency among classes.

Note that for a class of patients, their admission hours h_{adm} are ordered and thus cannot be iid. Although the LOS and h_{dis} of these patients are two independent iid sequences, it follows from (1) that their service times are no longer exogenous variables and are not iid.

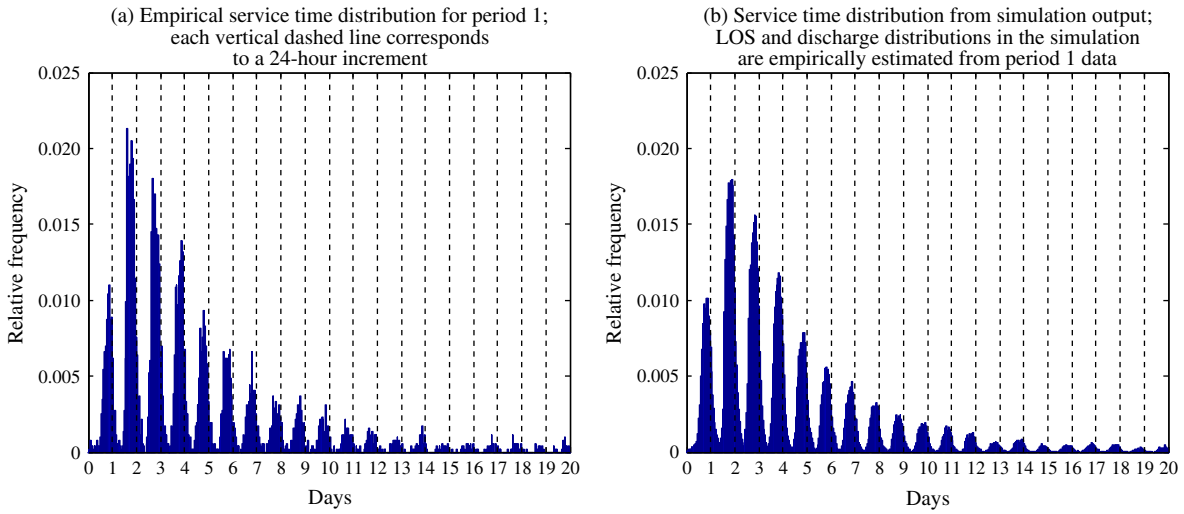
3.2.2. Separation of Time Scales. In the service time model (1), we use LOS to capture the number of nights that a patient needs to spend in the hospital as a consequence of medical conditions. We use the other two terms to capture the extra amount of time that is caused by operational factors. In particular, the discharge hour h_{dis} depends on discharge patterns that are mainly the results of schedules and behaviors of medical staff. The way we model the service time allows us to evaluate a variety of policies that may affect the two parts of the service time (LOS versus $(h_{\text{dis}} - h_{\text{adm}})$) jointly or separately. For example, the early discharge policy implemented at NUH aims to reduce the operational bottlenecks and move the discharge hour h_{dis} to an earlier time of the day without affecting the patient's medical conditions (LOS), whereas expanding the capacity at a nursing home or a step-down care facility to ensure timely discharge of patients in need of long-term care will mainly affect the LOS term (Borghans et al. 2008). In §6, we use simulation to gain managerial insights into the impact of early discharge and other policies on the waiting time performance.

Moreover, this service time model captures an interesting phenomenon, the separation of time scales: the LOS is in the order of days, whereas $(h_{\text{dis}} - h_{\text{adm}})$ is in the order of hours. Indeed, we can observe these two time scales from Figure 5(a), which plots the empirical service time distribution at hourly resolution. On the one hand, the distribution peaks at integer values representing $1, 2, 3, \ldots$ days, which is captured by the LOS. On the other hand, the sample points distribute around the integers mostly within the range of a few hours, which is captured by the term $(h_{\text{dis}} - h_{\text{adm}})$. Figure 5(b) illustrates that our proposed service time model (1) can produce the distributions that resemble empirical distributions. The two time scales (hour versus day) have been discovered in other studies of hospital operations (Armony et al. 2015, Mandelbaum et al. 2012, Ramakrishnan et al. 2005) and appointment scheduling (Zacharias and Armony 2015).

3.3. Critical Feature 2: Bed Assignment with Overflow

In this section, we spell out the details for bed assignment under a specified service policy. In particular, we described the overflow mechanism in our model.

Figure 5 **(Color online) Service Time Distributions, at Hourly Resolution, for General Medicine Patients Admitted in the Afternoon**



When a patient makes a bed request, if a primary bed is available, that bed is assigned to the patient. When more than one primary pool has such a bed, a priority policy included in the service policy is used to decide from which primary pool to select.

If no primary bed is available at the bed-request time, the patient waits in a buffer and is assigned with an overflow trigger time T. The trigger time T may depend on the bed-request time, the admission source, and the specialty of the patient. An overflow policy dictates the choice of T. The patient waits for a primary bed before the waiting time reaches T. After that, the patient can be assigned to either a primary bed or an overflow bed, whichever becomes available first.

Note that patients can be overflowed to a nonprimary server pool only if the waiting time exceeds the trigger time T. When T is not zero, a bed can be idle even if a patient from a nonprimary specialty has been waiting. Therefore, in our model the overflow policies are in general idling, which is different from the nonidling policies employed in many existing queueing models (Kumar 1993).

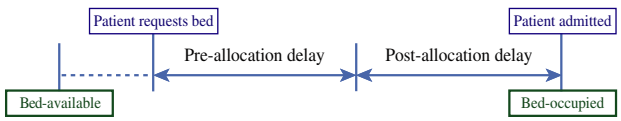
Overflow is an important measure for hospitals to balance the random demand and supply of different beds and to admit patients in a reasonably short time, given that it is difficult to adjust bed capacity among various specialties and wards in a short time window. (This is in contrast to call center operations, where the agents can be added or removed in a matter of hours.) NUH data show that the *partial* resource sharing from such overflow provides enough flexibility for hospitals to run in the quality-and-efficiency driven (QED) regime, in which the average patient waiting time (in the order of a few hours) is a small fraction of the average service time (in the order of days) and the bed utilization is high, say, more than 90%. A QED regime is usually gained by pooling a large number of servers (e.g., hundreds of beds) working in parallel and is difficult to achieve for a small number of servers (e.g., 30 beds in a ward).

3.4. Critical Feature 3: Pre- and Post-Allocation Delays

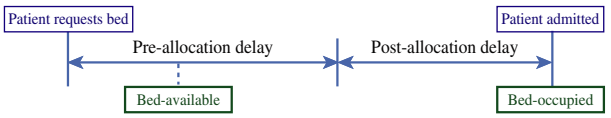
In reality, to admit a patient into an inpatient ward, one needs not only an available inpatient bed but also many other secondary resources to coordinate the transfer process from the ED to the ward (e.g., ED and ward nurses). Constraints on these secondary resources will delay the transfer process even if a primary bed is readily available. Thus, we explicitly incorporate the operational delays that are caused by these secondary resource constraints in the model. To be more specific, even if a primary bed is available for each patient upon arrival, each patient has to experience a pre-allocation delay first, and then a post-allocation delay before being admitted to the bed. In §3.4.1, we first describe the process flow from a patient's bed request to admission to a bed in our model, and then in §3.4.2, we explain the rationale of modeling the two allocation delays. Figure 6 illustrates the process with two allocation delays under various scenarios to be specified in §3.4.1.

3.4.1. Patient Flow from Bed Request to Admission. In our model, when a patient makes a bed request, we assume two bed-allocation modes: normal allocation and forward allocation. The two modes differ from each other with respect to when the patient starts to experience a pre-allocation delay. In a normal allocation, the patient starts to experience a pre-allocation delay immediately at the bed-request time if a primary bed is available at that time (Figure 6, case A). If no primary bed is available, the patient waits in a buffer for a bed. When a bed becomes available and is assigned to that patient, following the bed assignment policy described in §3.3, the patient starts to experience a pre-allocation delay

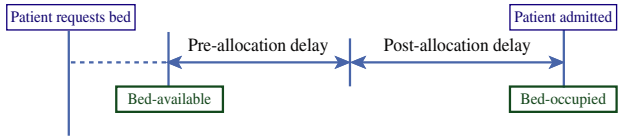
Figure 6 (Color online) Pre- and Post-Allocation Delays Under Different Scenarios



Case A. Normal allocation: A bed is available before patient requests bed



Case C. Forward allocation: A bed is available before pre-allocation delay expires



Case B. Normal allocation: A bed is available after patient requests bed

Case D. Forward allocation: A bed is available after pre-allocation delay expires

(Figure 6, case B). In a normal allocation, this pre-allocation delay always begins at or after the bed-available time.

A forward allocation is used only when there is no primary bed available at the patient's bed-request time (Figure 6, cases C and D). The patient starts to experience a pre-allocation delay immediately at the bed-request time. In other words, a pre-allocation delay always begins before a bed becomes available in the model. Therefore, sometimes a bed may still be unavailable when the patient finishes the pre-allocation delay stage.

In general, a patient starts to experience a post-allocation delay when the pre-allocation delay expires. The only exception is when the forward allocation mode is used and a patient finishes experiencing a pre-allocation delay but a bed is still unavailable (Figure 6, case D). In this case, the patient waits until a bed becomes available, and a post-allocation delay starts at the bed-available time. When the post-allocation delay expires, the patient is admitted into the bed, completing the bed-request process.

We assume that a bed-request at time t, if there is no primary bed available, has probability $p(t)$ to be a normal allocation and probability $1-p(t)$ to be a forward allocation. We assume that the pre- and post-allocation delays are independent random variables following certain continuous distributions. The means of the distributions can be time dependent, depending on when the patient requests a bed and starts to experience the allocation delays.

3.4.2. Rationale for Modeling and Other Remarks. In practice, allocating a bed to an incoming patient is a process. We use the pre-allocation delay to model the time needed for the bed management unit (BMU) to search and negotiate a bed for a patient from an appropriate ward. The start and end points of the pre-allocation delay correspond to when a BMU agent starts and finishes the bed-allocation process, respectively. At the end of the bed-allocation process, a bed is allocated to the patient and NUH registers this time as the allocation-completion time. However, the allocation-completion time does not necessarily correspond to the time when a bed is assigned to a patient in our model; the bed assignment in our model is specified in §3.3 and always happens at a patient's bed-request time, overflow trigger time, or discharge time. For example, if a primary bed is available upon a patient's bed request, the bed assignment is instantaneously done in our model before the patient starts to experience the pre-allocation delay.

We use the post-allocation delay to model the delay after a bed is allocated and available to use for an incoming patient. These delays include the time needed to discharge the patient from the ED or a nongeneral ward and transport that patient to a GW. Thus, the start point of the post-allocation delay corresponds to the allocation-completion time or the bed-available time, whichever is later, and the end point corresponds to the patient's admission time in practice.

Among the time stamps mentioned in the previous two paragraphs, NUH does not record when the bed-allocation process starts. According to our interviews and empirical analysis at NUH (Shi et al. 2014), BMU agents normally wait until a bed becomes available before starting the bed-allocation process (which is close to the normal-allocation mode), or sometimes they can forward-allocate a bed based on the planned discharge information (which is close to the forward-allocation mode). We use the normal- and forward-allocation modes to approximate this reality. Note that the actual allocation mode in practice may be neither normal nor forward as in the model, since the starting time of the actual bed-allocation process may be somewhere between the bed-request and bed-available times. Thus, an alternative setting is to randomly assign this starting time to occur between the bed-request and bed-available times following a certain distribution. We leave this extension to a future study.

3.5. Service Policies

A service policy governs all of the decisions regarding bed assignments at various decision time points. It has four components: (i) how to pick a bed from a primary pool upon an arrival, (ii) how to pick a bed from a nonprimary pool when a patient's overflow trigger time is reached, (iii) how to set an overflow trigger time, and (iv) how to pick a waiting patient from a group of eligible patients upon the departure of another patient. We elaborate each component below.

Component (i) specifies the priority of primary pools for each of the specialties having more than one primary pool. In general, dedicated pools (which serve only one specialty) have higher priorities than shared pools (which serve multiple specialties). Therefore, when seeking a primary bed for a patient, we start from the dedicated pools. If there is no dedicated bed free, we then search in shared pools.

Component (ii) specifies the priority of nonprimary pools for overflowing patients. The priority depends on the specialty of the patient to be overflowed. In general, pools that serve similar specialties have high priority. Shared pools have higher priority than dedicated pools. Both components (i) and (ii) need to be estimated based on the actual configuration in the particular hospital being modeled.

Section 3.3 has introduced an overflow mechanism in our model. Component (iii) sets the overflow trigger time T for patients who have to wait because of the unavailability of primary beds upon their arrivals.

When a patient's waiting time reaches the trigger time T, component (ii) is used to search for a non-primary bed for that patient. Different hospitals may adopt different overflow policies, and we will specify the time-dependent dynamic overflow policy adopted at NUH in §4.6.

Component (iv) is a patient priority list, which is used when a bed becomes available and needs to be assigned to one of the eligible patients. The eligible patients consist of both the primary patients and the overflow patients whose waiting times are greater than their overflow trigger times. Again, this component needs to be estimated according to each hospital's own situation. Generally speaking, patients who have waited longer than their overflow trigger times have a higher priority than those who have not.

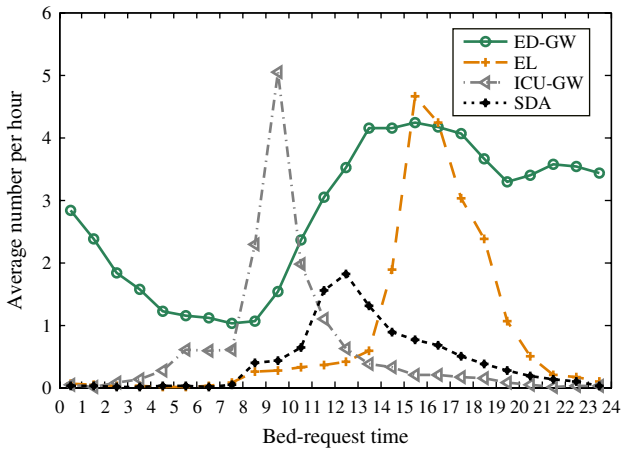
4. Populated Stochastic Model Using NUH Data

Based on the empirical study at NUH (Shi et al. 2014), we populate the proposed stochastic network model, which we hereinafter refer to as the "NUH model". In this section, we discuss how we empirically estimate all the necessary input for the NUH model. Unless stated otherwise, we always use period 1 data to estimate the input, and the resulting NUH model is called the "baseline scenario." Section 4.1 introduces the arrival processes for the four admission sources. Section 4.2 describes the server pool setting and the service policy. Section 4.3 discusses the adjustments we have made to incorporate patient transfers between GWs and ICU-type wards. Section 4.4 presents the empirical LOS and discharge distributions, and §4.5 introduces classification of patients based on the observations of LOS distributions. Sections 4.6 and 4.7 illustrate a dynamic overflow policy and time-varying allocation delays for the NUH model, respectively. Some remarks on the empirical study are in §4.8.

4.1. Arrivals

4.1.1. Time-Varying Arrival Rates. As shown in Figure 4, patient arrivals in our model derive from four sources. For each source, the arrival rate depends on the time of day. For ED-GW, ICU-GW, and SDA patients, we use their empirical, hourly bed-request rates as their arrival rates in the NUH model. The arrivals of EL patients are prescheduled. NUH has their admission times but lacks meaningful records of bed-request times. Thus, we use their empirical, hourly admission rates as their arrival rates in the NUH model. We assign EL patients the highest priority and set their allocation delays to be zero. In this way, the waiting times of EL patients in the NUH model are negligible, and hence their admission times are close to their bed-request times. Figure 7 shows the estimated hourly arrival rates for the four admission sources in the course of a day.

Figure 7 **(Color online) Hourly Arrival Rate for Each Admission Source (Estimated from Period 1 Data)**



Note. The daily arrival rate of each source is close to its daily admission rate as shown in Figure 3(a), except for the ICU-GW source, since readmitted patients are included here.

4.1.2. Arrival Processes. *A Nonhomogeneous Poisson Process for ED-GW Patients.* For the empirical bed-request process of ED-GW patients, we conducted a detailed study to test the assumption that it is a nonhomogeneous Poisson process. Following a statistical procedure proposed in Brown et al. (2005), we perform 30 tests, one for each month in the two periods, on the empirical bed-request times. Among the 30 tests, 24 of them do not reject the hypothesis that the bed-request process of ED-GW patients follows a nonhomogeneous Poisson process with piecewise-constant arrival rates. Therefore, it is reasonable to assume that the bed-request process for ED-GW patients is nonhomogeneous Poisson. However, we find that the bed-request process is not a periodic Poisson process with either one day or one week as a period. In particular, the empirical coefficient of variation (CV) of the daily arrival rate for each day of week is much higher than one, the theoretical CV under the Poisson assumption. We conjecture that the high variability comes from the seasonality of bed requests and the overall increasing trend in the bed demand (Shi et al. 2014). In the NUH model, we assume that the ED-GW patient's arrival process is nonhomogeneous Poisson. We further assume that it is periodic with one day as a period. The arrival rate function of the periodic Poisson process is constant in each hour and is plotted as the solid curve in Figure 7. Note that setting one week as a period is another reasonable choice, and we discuss this extension as a future study to capture the day-of-week phenomenon in §7.1.

A Non-Poisson Arrival Process Model for Other Sources. The number of EL admissions each day is prescheduled at NUH. The bed requests of ICU-GW

or SDA patients are departures from the ICU-type or SDA wards, and their volumes are in a way also prescheduled on a daily basis: ICU physicians determine the number of patients to be transferred to GWs after the morning rounds each day, and then ICU nurses submit the bed requests for these ICU-GW patients; similar to the EL patients, the SDA surgery patients each day are scheduled in advance, and the SDA nurses submit bed requests after the SDA patients finish receiving surgeries on that day. Based on this observation, we propose a non-Poisson arrival model for EL, ICU-GW, and SDA patients. We first generate a total number of A_k^j arrivals (to arrive on day k) from admission source j ($j = 1, 2, 3$, denoting EL, ICU-GW, and SDA, respectively) at the beginning of day k ($k = 0, 1, \ldots$), where the value of A_k^j is randomly generated from the empirical distribution of the daily number of bed requests for $j = 2, 3$ or daily admissions for $j = 1$. We then randomly assign the arrival times of A_k^j arrivals according to order statistics that draw from the empirical distribution of bed-request (or admission) times of source j. These distributions can be estimated from the arrival rate curves in Figure 7. Note that if the daily number of arrivals follows a Poisson distribution, the generated process is in fact a nonhomogeneous Poisson process with one day as the period (Lewis and Shedler 1978).

4.2. Server Pools and Service Policy

In the NUH model, there are 15 server pools. Table 1, which is constructed based on the period 1 data, lists the number of servers and the primary specialties for each server pool. Note that the number of servers in this table does not fully match the actual bed capacity in period 1. We have made some adjustments to account for the fact that the bed capacity had been increasing in period 1 and some wards had disproportionally high overflow proportions. Details of these adjustments and the three overflow wards in Table 1 (with pool ID 12, 13, and 14) are specified in §4.1 of the online supplement (available at http://dx.doi.org/10.1287/mnsc.2014.2112).

Table 1 Server Pool Index, Primary Specialty, and Number of Servers

Pool ID	Primary specialty	No. of servers
0	Gen Med, Respi	41
1	Gen Med, Neuro	40
2	Renal	33
3	Neuro	12
4	Gastro-Endo	39
5	Surg	42
6	Cardio	40
7	Ortho	50
8	Onco	43
9	Respi, Surg	25
10	Surg, Ortho	38
11	Surg, Cardio	30
12	Overflow ward I	39
13	Overflow ward II	43
14	Overflow ward III	48
Total		563

Table 2 Priority of Primary and Overflow Pools

Specialty	Primary	Overflow
Surg	5, 10, 11, 9	14, 12, 13, 7, 4, 1, 0, 2, 3
Card	6, 11	13, 14, 12, 4, 10
Gen Med	0, 1	14, 13, 4, 2, 3, 9, 10, 12, 8, 7, 11, 5, 6
Ortho	7, 10	12, 5, 14, 13, 4, 1, 2
Gastro-Endo	4	14, 13, 1, 0
Onco	8	13, 14, 1
Neuro	3, 1	14, 13, 4, 2, 0, 9, 10, 8, 7, 11
Renal	2	1, 4
Respi	9, 0	14, 13, 1, 4, 2, 3, 10, 8, 7, 11, 5

Note. Pool numbers are ordered in decreasing priority.

The service policy is built based on NUH's internal guidelines (National University Hospital 2011) and our empirical observations. Specifically, Table 2 gives the priority table for components (i) and (ii) of the service policy discussed in §3.5. Component (iii), the overflow policy, will be elaborated in §4.6.

The priority list of component (iv) is given below. First, patients who have waited longer than their overflow trigger times have a higher priority than those who have not. This is aligned with NUH's goal of improving the 6-hour service level. Second, among the patients waiting longer than their overflow trigger times, those from the primary specialties have a higher priority than the ones from overflow specialties. Third, among patients from the same specialty, the ED-GW patients have a higher priority than ICU-GW and SDA patients, whereas ICU-GW and SDA have the same priority. This is based on the empirical observation that, at NUH, ICU-GW and SDA patients have a much longer average waiting time than ED-GW patients (see §2.1). Also see Powell et al. (2011) for a similar priority setting. Moreover, our model assumes that EL patients have the highest priority among all admission sources to account for using admission times as a proxy for bed-request times; see reasons in §4.1. Fourth, when patients are waiting in multiple buffers with the same priority or in a single buffer, we choose the patient with the longest waiting time.

4.3. Modeling Patient Transfers Between ICU-Type Wards and GWs

The empirical study at NUH (Shi et al. 2014) shows that a patient from any admission source can be transferred between a GW and an ICU-type ward, possibly multiple times, after the initial admission to the GW. To incorporate the transfer patient flows between the GWs and the ICU-type wards, an obvious choice is to model explicitly the ICU-type wards as a separate

group of server pools. However, the time stamps in our data sets do not have high-enough resolution for us to conduct a detailed empirical analysis for patient flows within the ICU-type wards. Thus, we are unable to adopt this approach in the present study. Alternatively, we stay within the single-pass stochastic processing network framework and make the following adjustments to incorporate the transfer patient flows in the NUH model.

We choose to model those real patients who are admitted from ED-GW or EL sources at NUH and have been transferred once or twice between GWs and ICU-type wards after their initial admissions (i.e., from GW to ICU or from GW to ICU then back to GW, respectively). We do not model (i) the real patients who are initially admitted from ICU-GW or SDA sources and have been transferred or (ii) ED-GW or EL patients who have transferred more than two times. We exclude them because the volume of these patients is small. For the chosen real patients, we use two steps to model their stays in the GWs.

1. In step 1, we determine an arriving ED-GW or EL patient (in the model) to be a nontransfer type or a transfer type patient upon arrival according to a certain Bernoulli distribution. The parameters of these distributions will be specified in §4.5. The transfer type patient models the initial admission to a GW of a real patient who will later be transferred to an ICU-type ward. The reason we need to differentiate between nontransfer and transfer type patients in the model is that their LOS and discharge time distributions are significantly different, even if they are from the same admission source and specialty; see details in §4.4 below.

2. NUH data shows that 71% of the transfer patients we choose to model will be transferred back to a GW after their first transfer to an ICU-type ward. Thus, they will be admitted to a GW for the second time. In step 2, we use a separate patient class to model this second admission to GWs. We further assume that this separate patient class is ICU-GW sourced in the model, because our empirical analysis shows that the characteristics of the second admissions to GWs, including the admission time distributions, are close to those of the initial admissions from ICU to GWs (the latter corresponds to original ICU-GW patients in our model). We call the class of patients that is artificially created to accommodate the second admission to GWs the "readmitted" ICU-GW patients, and the other ICU-GW patients the "newly admitted" ICU-GW patients. Again, the reason we need to differentiate the two classes of ICU-GW patients is that their LOS distributions are significantly different. When we generate an arrival from the ICU-GW source, we determine whether the arriving patient is newly admitted or readmitted according to a certain Bernoulli distribution. Patient class in the NUH model will be further specified in §4.5.

4.4. Length of Stay and Discharge Distributions

4.4.1. Nontransfer Patients. Table 3 lists the empirically estimated mean and standard deviation of LOS for nontransfer type patients (or newly admitted patients for the ICU-GW source) from different admission sources and specialties in the NUH model. Transfer type and readmitted patients in the model have a different set of LOS distributions, and we discuss estimating their LOS distributions in §4.4.2.

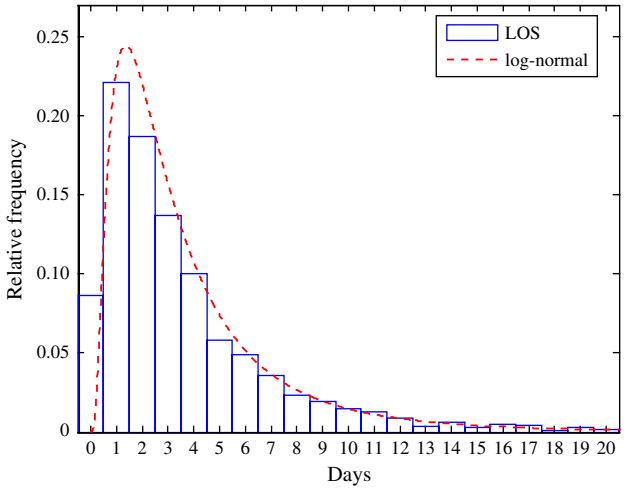
Not surprisingly, admission source and specialty affect a patient's LOS. We want to emphasize that LOS distributions are also admission-period dependent for ED-GW patients. Table 3 shows that for each specialty of ED-GW patients, a before-noon admission (i.e., A.M.) patient on average spends one day less than an after-noon admission (i.e., P.M.) patient. We speculate the reason might be that the rest of the admission day can be used for further medical diagnosis for A.M. patients, but not for P.M. patients. For patients from the other three admission sources, we do not assume their LOS to be admission-period dependent because there are very few A.M. patients from these sources.

Table 3 Average LOS (in Days) for Nontransfer Patients in Each Specialty from Different Admission Sources

Cluster	ED-GW(AM)	ED-GW(PM)	EL	ICU-GW	SDA
Surg	2.36 (2.93)	3.27 (3.43)	4.55 (6.55)	9.58 (12.60)	2.59 (4.72)
Cardio	2.95 (3.75)	3.83 (3.93)	4.15 (5.08)	5.22 (6.78)	2.55 (3.38)
Gen Med	3.94 (4.76)	5.25 (5.87)	5.32 (5.79)	10.43 (18.43)	3.17 (2.62)
Ortho	5.45 (8.22)	6.04 (7.04)	6.27 (6.19)	10.82 (13.32)	3.41 (4.32)
Gastro-Endo	3.32 (3.91)	4.48 (4.47)	3.70 (4.39)	8.33 (12.25)	3.24 (3.99)
Onco	5.93 (7.58)	7.03 (7.14)	6.45 (7.95)	8.62 (9.02)	4.10 (4.18)
Neuro	3.23 (5.22)	4.07 (4.69)	4.06 (4.69)	7.56 (7.67)	2.59 (2.40)
Renal	5.75 (6.55)	6.51 (6.90)	5.70 (6.20)	10.22 (12.91)	2.08 (1.16)
Respi	3.21 (5.10)	4.29 (4.26)	4.45 (6.27)	7.86 (10.71)	2.33 (3.33)
All	3.70 (5.25)	4.78 (5.45)	5.17 (6.47)	7.59 (10.82)	2.84 (4.29)

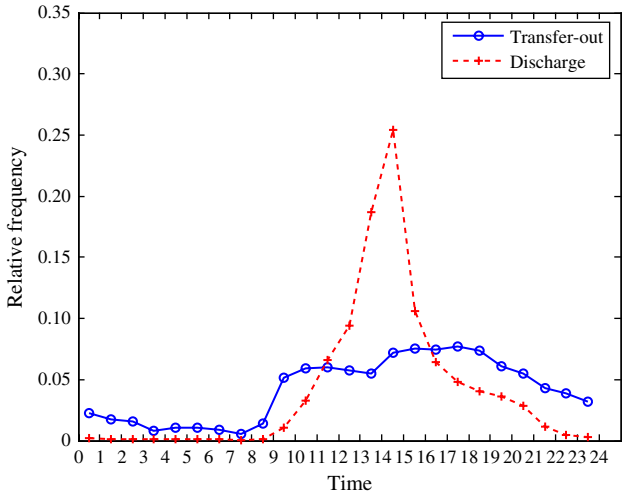
Notes. Period 1 data are used, and the number in the parentheses is the corresponding standard deviation. Here, we list the standard deviation instead of the confidence interval to help readers to fit distributions for LOS.

Figure 8 **(Color online) LOS of ED-GW Patients from General Medicine**



Notes. Only nontransfer A.M. patients in period 1 are included. The LOS distribution can be fitted with a log-normal distribution (mean = 3.94, SD = 4).

Figure 9 **(Color online) Discharge Distributions for Transfer Patients from ED-GW and EL Sources and for Nontransfer Patients**



Note. The solid curve is estimated from combining the data from periods 1 and 2, and the dashed curve is estimated from period 1 data.

In the NUH model, we use empirical LOS and discharge distributions estimated from the data. The discharge distributions in the two periods are plotted in Figure 2(a). Figure 8 illustrates the LOS distribution of general medicine ED-GW patients who are admitted before noon. Plots of other LOS distributions have similar shapes.

4.4.2. Transfer Patients. Based on the description in §4.3, we use the first-visit LOS and transfer-out times of the corresponding real patients to estimate the LOS distributions and discharge distributions for the transfer type ED-GW or EL patients, respectively. We use the second-visit LOS of the real patients who transferred twice to estimate the LOS distributions for the readmitted ICU-GW patients. The discharge distribution of the readmitted ICU-GW patients is the same as the one for the nontransfer patients (as shown in Figure 2(a)). Figure 9 plots the discharge (transfer-out) distribution for all the transfer type ED-GW and EL patients. We do not observe a significant difference between the two periods.

4.5. Patient Class

Patients belonging to the same class are homogeneous, having the same LOS and discharge distributions. The empirical evidence in §4.4 has shown that the LOS distributions depend on admission source, medical specialty, admission period (for ED-GW patients), and whether patients are transferred or not. Based on §§4.3 and 4.4, we proceed in the following steps to determine a patient's class in the NUH model:

1. When an arrival from one of the four admission sources occurs, we assign this patient to one of the nine medical specialties, following an empirical distribution that depends on both the bed-request hour and admission source. Figure 3(b) plots the daily distributions of specialties and admission sources. After assigning the specialty, the service priority of the patient is determined. The following two steps make sure that the LOS and discharge distributions are the same within a class.

2. Next, we determine (i) whether an ED-GW or EL patient is a nontransfer or a transfer patient or (ii) whether an ICU-GW patient is newly admitted or readmitted following a Bernoulli distribution that depends on the specialty. The parameters for these Bernoulli distributions are empirically estimated based on the relations between the patients in the model and real patients who have transferred (see §4.3) and are listed in Table 4.

3. Finally, at an ED-GW patient's admission time, we determine that patient's admission period (A.M. or P.M.). By now, the patient's class is fully determined.

Table 4 **Estimated Value for the Parameter p of the Bernoulli Distribution to Determine Patient Classes**

p	Surg (%)	Cardio (%)	Med (%)	Ortho (%)	Onco (%)
ED-GW	4.58	11.52	4.78	9.42	5.69
EL	23.46	39.95	4.53	17.04	6.01
ICU-GW	45.10	43.86	16.98	79.69	39.86

Notes. For ED-GW and EL patient types, p represents the probability of being a transfer patient; for ICU-GW, p represents the probability of being a readmitted patient. Parameters for specialties belonging to the medicine cluster (Gen Med, Gastro-Endo, Neuro, Renal, Respi) are estimated together due to the limited number of data points, and we use Med to represent this group.

4.6. A Dynamic Overflow Policy

At NUH, there is a general guideline (National University Hospital 2011) on when and how to overflow a patient. Consistent with this guideline, empirical evidence (Teow et al. 2012) suggests that the hospital overflows patients more aggressively during late night and early morning (before 7 A.M.). That is, NUH will overflow a patient almost immediately upon finding that no primary bed is available. The reason is that few discharges happen in this time period, so there is little chance that a primary bed will become available in the next few hours. Thus, there is no need to let the patient wait for another hour. In contrast, during other times, the hospital tends to be more conservative and allows a patient to wait some time before overflow in anticipation that a primary bed may become available soon. In this way, NUH has better control on the overflow proportion, another important performance metric being monitored (see §2.3.2). The preceding discussion suggests that the trigger time T should depend on the bed-request time. It is reasonable to assume that T is low when a bed request occurs during late night or early morning, and high during other times.

Based on these observations, we use a simple dynamic overflow policy in the NUH model: when a patient requests a bed from 7 A.M. to 7 P.M., the overflow trigger time T is set to be $t_2 = 5.0$ hours; for bed requests in all other time periods, T is set to be $t_1 = 0.2$ hour. We choose 7 A.M. and 7 P.M. as the starting and ending points, respectively, to adopt the long overflow trigger time. This choice is based on observations from Teow et al. (2012) and the practice at NUH. The night shift period at NUH is from 7 P.M. to 7 A.M. the next day. A nurse manager is in charge of dealing with all bed requests in this period and has the authority to overflow patients without negotiation. The values of t_1 and t_2 are obtained through trial and error so that the simulation output curves are as close to the empirical curves as possible. It is important to note that overflow decisions are very complicated (Teow et al. 2012), sometimes subjective, in practice. There is no data available for us to get an accurate estimation of the overflow trigger time. Thus, our proposed dynamic policy is only an approximation of the reality, and the two values t_1 and t_2 are not directly estimated from data. Readers are referred to the online supplement for our sensitivity analysis using alternative overflow policies.

4.7. Pre- and Post-Allocation Delays

In this section, we focus on estimating allocation delays for ED-GW patients. We first explain how to model allocation delays for other patients. We assume the allocation delays of the EL patients to be zero in the model, having explained the rationale of doing so in §4.1. For ICU-GW and SDA patients, we do not have good time stamps to estimate their pre- and post-allocation delays reliably. We simply assume that their allocation delays follow the same distributions as the ones used to generate the allocation delays for ED-GW patients. Sensitivity analysis shows that a moderate amount of change to the allocation delay distributions of ICU-GW and SDA patients will not affect the overall performance of ED-GW patients.

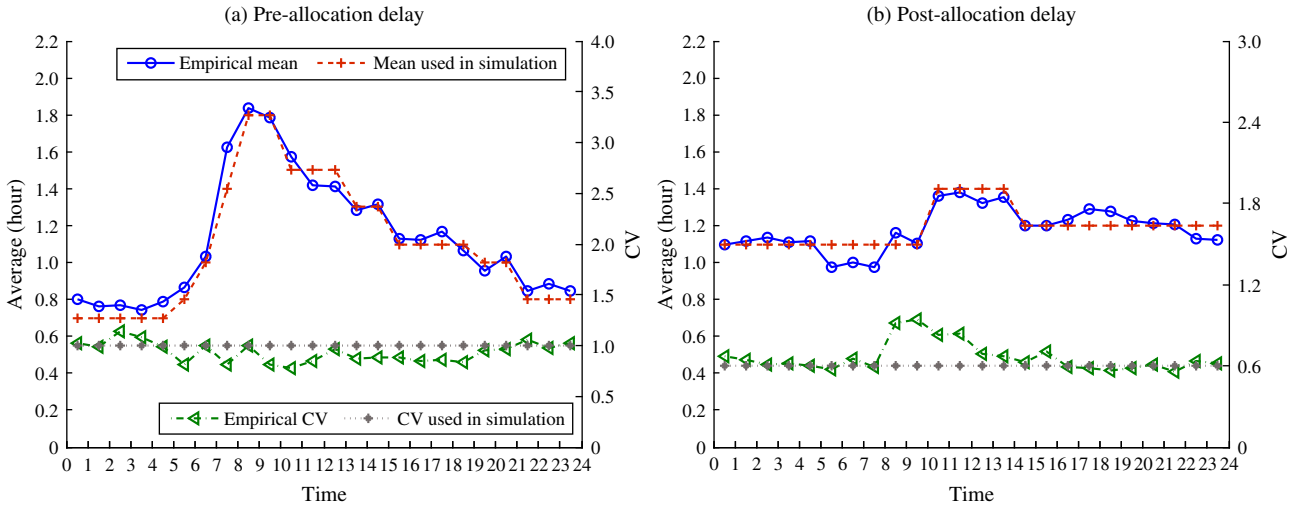
4.7.1. Distributions of the Time-Dependent Allocation Delays. In the NUH data set, at the bed-request time of an ED-GW patient either (i) the allocated bed is already available for the patient or (ii) the bed is not available and is still occupied by another patient. Case (i) corresponds to case A in Figure 6, and we select a subset of case (i) patients in the data set to estimate the pre-allocation delay distribution. The subset consists of case (i) patients whose allocated beds are from their primary wards. By selecting this group of patients, we try to minimize the influence of bed shortage and specialty mismatch on pre-allocation delay so that our estimation can reflect the minimum time needed for BMU agents to allocate a bed. For the post-allocation delay, there is no such influence and we include all ED-GW patients to estimate its distribution.

From §3.4.2, we know how to estimate the allocation delays empirically from the time stamps recorded in NUH data. Subgroups are created to account for the time-dependent feature of allocation delays, which we will explain in the following paragraphs. The histograms and distributional fitting results suggest that using a log-normal distribution is a good starting point for modeling each of the allocation delays. Thus, our model assumes the pre- or post-allocation delay initiated within each hour of a day to be an iid random variable that follows a log-normal distribution. The mean and CV of the log-normal distribution depends on the initiation hour (i.e., the hour when the allocation delay starts).

Figures 10(a) and 10(b) plot the empirical estimates of the mean and CV for the pre- and post-allocation delays, respectively. In our baseline scenario, we use the two dashed curves denoted with a plus sign as the inputs for the time-dependent mean and CV for each allocation delay, respectively. These two curves are slightly smoother than (but still within the 95% confidence intervals of) the corresponding empirical curves, which have random noise since the sample sizes in certain time intervals are small, particularly between 8 A.M. and 10 A.M.

The empirical curves in Figures 10(a) and 10(b) clearly demonstrate a time-dependent feature of both allocation delays. The average delays are longer if the delay initiation time is in the morning, especially for the pre-allocation delay. The longer pre-allocation

Figure 10 **(Color online) Mean and CV of Estimated Pre- and Post-Allocation Delays with Respect to the Delay Initiation Hour**



Notes. The left vertical axis is for the average; the right vertical axis is for the CV. The scale of the right vertical axis is deliberately chosen to be large, so that the four curves are not crossed over.

delay in the morning may stem mainly from the ward side. At NUH the ward physicians and nurses are busy with morning rounds, and therefore it may take the BMU longer time to search and negotiate for beds. The longer post-allocation delay in the morning may stem mainly from the ED side. The ED at NUH is usually congested in the late mornings, so it is likely that ED physicians and nurses are busy with newly arrived patients and have less time to discharge and transfer admitted patients to wards.

4.7.2. Estimating the Normal Allocation Probabilities $p(t)$. Recall from §3.4 that in the model, when a patient makes a bed request at time t and there is no primary bed available, we assume with probability $p(t)$ that the allocation for the patient is a normal allocation, meaning this patient will wait until a bed is available before starting to experience the pre-allocation delay. Unfortunately, the NUH data set does not have accurate time stamps to allow us estimate $p(t)$ reliably. In our baseline scenario, we choose

$$p(t) = \begin{cases} 0 & h(t) \in [0, 6), \\ 0.25 & h(t) \in [6, 8), \\ 1 & h(t) \in [8, 12), \\ 0.75 & h(t) \in [12, 14), \\ 0.5 & h(t) \in [14, 20), \\ 0 & h(t) \in [20, 24), \end{cases} \tag{2}$$

where $h(t)$ stands for the hour of the day of the bed-request time t. The choice of $p(t)$ is based on the current practice at NUH and empirical estimation of the proportion of patients whose bed-allocation process approximately corresponds to the normal-allocation mode in the model. Section 4.2 of the online supplement discusses the details of estimating $p(t)$ in different time intervals. We realize that, despite our best efforts, our choice of $p(t)$ using (2) is still ad hoc. We report a sensitivity analysis of the choice of $p(t)$ in §6.4.

4.8. More Empirical Evidence

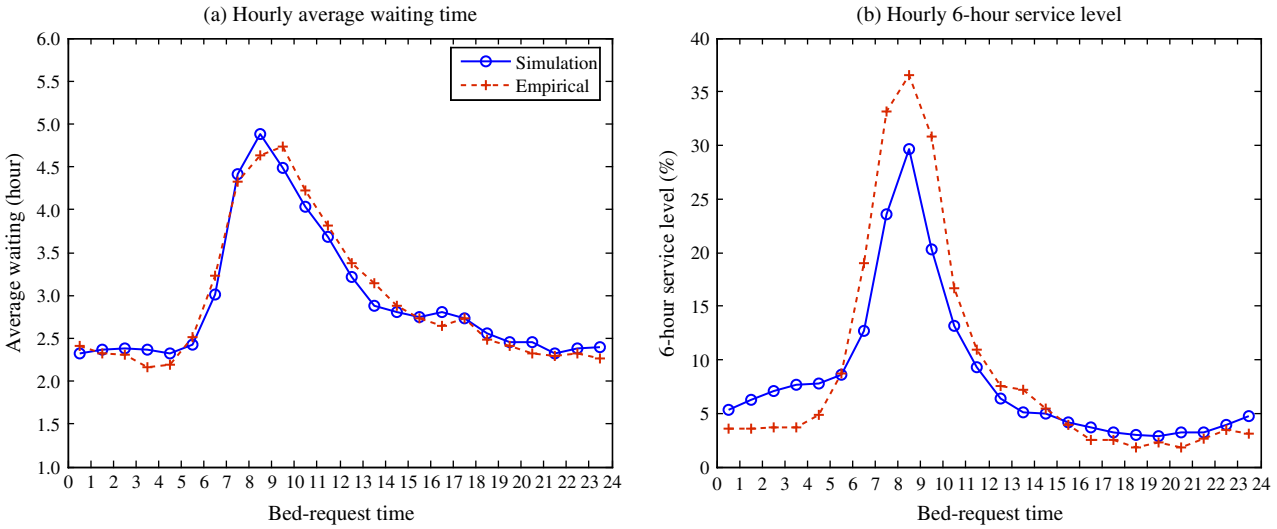
In the interest of space, many of the empirical observations, distributional fitting, statistical results, and tables and plots cannot be displayed in this paper. We refer the readers to the companion paper (Shi et al. 2014) for the details. Most of these results are also available in Shi (2013), and we specify some of the section numbers in Shi (2013) here for readers' convenience. Section 2.4 in Shi (2013) presents statistics and tests results for the arrival processes. Sections 2.5 and 2.6 document the empirical distributions for LOS and service times. Section 2.7 illustrates the bed-allocation and admission processes at NUH and more evidence to support our modeling of the allocation delays. Section 2.8 contains the statistics for transfer patients and plots of their LOS distributions.

5. Verification of the Populated NUH Model

Recall that the populated NUH model, using the input described in §§4.1–4.7, is referred to as the "baseline scenario." In §5.1, we first show that the simulation output from the baseline scenario matches several key empirical performance measures. Then, in §5.2, we show that the simulation output from each model that misses one of the three critical features introduced in §3 cannot replicate the empirical performance measures.

To implement these models, we wrote a simulation code in C++ language. For each simulation run, we start from an empty system and simulate for a total

Figure 11 **(Color online) Baseline Simulation Output Compares with Empirical Estimates: Hourly Average Waiting Time and 6-Hour Service Level (Period 1)**



of 10^6 days. We then divide the simulation output into 10 batches. The performance measures are calculated by averaging the last 9 batches, with the first batch discarded to eliminate transient effects. Unless otherwise specified, all simulation estimates in this paper are from simulation runs under this setting. The choice of the simulation setting is justified following standard techniques in the literature (Law and Kelton 2000). Note that in this and the next section, we rely on simulation to obtain the desired performance measures, because there is no existing analytical tool to analyze the proposed stochastic model either exactly or approximately. As mentioned in the introduction, this paper focuses on establishing a high-fidelity model that can capture the inpatient flow dynamics at hourly resolution. We discuss ongoing and future analytical research in §7.

5.1. The Baseline Scenario

Recall that the inputs for the baseline scenario are estimated from the NUH period 1 data. Thus, we compare the outputs from this scenario against the empirical performance in period 1 to verify the NUH model. From simulating the baseline scenario, we find that the daily average waiting time for all ED-GW patients is 2.82 hours and the daily 6-hour service level is 6.29%, close to what we observed empirically in period 1. Furthermore, Figure 11 shows that the simulation estimates approximately replicate the empirical estimates of the time-of-day (hourly) waiting time performance for all ED-GW patients. Table 5 compares the simulation estimates with the empirical estimates of the average waiting time and the 6-hour service level for each specialty. The relative difference in the mean waiting time is less than 5% for all nine specialties. Except for cardiology and renal, the absolute difference in the 6-hour service level is less than 0.8% (relative difference of less than 13%) for all other specialties. We can see that the waiting time statistics, even at the specialty level, can be approximately replicated by our simulation.

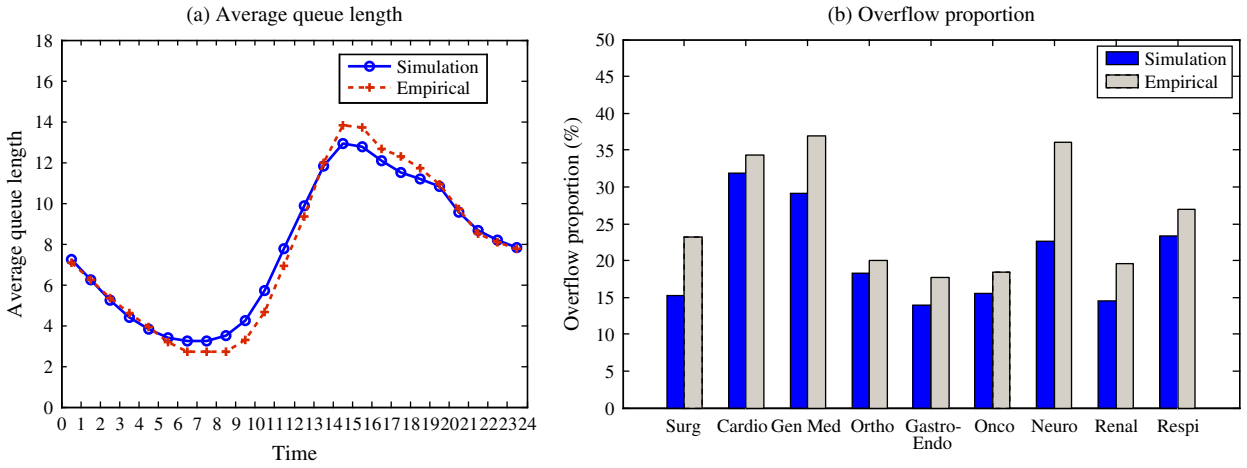
Besides the waiting time, we can also approximately replicate other key performance measures. The utilization rate is 89.2% from simulation, a little bit higher than the 88.0% empirical utilization in period 1. Figure 12(a) plots the hourly average queue length for all ED-GW patients for both simulation and empirical estimates. The average overflow proportion across all specialties is 21.70% under the baseline scenario, a slight underestimation of the empirical

Table 5 **Simulation and Empirical Estimates of Waiting Time Statistics for ED-GW Patients from Each Specialty**

	Average waiting time (hour)		6-hour service level (%)	
Specialty	Simulation	Empirical	Simulation	Empirical
Surg	2.64 ± 0.01	2.61 ± 0.05	4.85 ± 0.05	5.45 ± 0.57
Cardio	2.97 ± 0.01	3.08 ± 0.05	6.81 ± 0.06	8.36 ± 0.74
Gen Med	2.73 ± 0.01	2.64 ± 0.04	5.39 ± 0.05	4.79 ± 0.47
Ortho	2.73 ± 0.01	2.79 ± 0.05	5.22 ± 0.06	5.84 ± 0.68
Gastro-Endo	2.88 ± 0.01	2.97 ± 0.07	8.07 ± 0.07	7.64 ± 0.92
Onco	2.88 ± 0.01	2.96 ± 0.11	7.58 ± 0.06	8.15 ± 1.35
Neuro	2.84 ± 0.01	2.81 ± 0.07	6.49 ± 0.06	5.93 ± 0.9
Renal	3.23 ± 0.01	3.41 ± 0.09	10.45 ± 0.04	11.6 ± 1.31
Respi	2.82 ± 0.01	2.77 ± 0.09	6.25 ± 0.07	5.5 ± 1.14
All	2.82 ± 0.01	2.82 ± 0.02	6.29 ± 0.05	6.52 ± 0.26

Notes. The simulation estimates are from simulating the baseline scenario, and the empirical estimates are from period 1 data. The numbers $\bar{X} \pm \bar{H}$ in each entry denote the mean ($\bar{X}$) and the half-width of the 95% confidence interval ($\bar{H}$), respectively. The confidence intervals for the simulation output are calculated following the batch mean method (Law and Kelton 2000); the confidence intervals for the empirical statistics are calculated with the standard deviations and sample sizes from the actual data.

Figure 12 **(Color online) Baseline Simulation Output Compares with Empirical Estimates: Hourly Average Queue Length and Overflow Proportion (Period 1)**



estimates from period 1 data (26.95%). Figure 12(b) compares the overflow proportion from simulation and empirical estimates for each specialty. Readers are referred to §5.1 of the online supplement for more discussion on model limitations that lead to the underestimation of the overflow proportions.

In obtaining our baseline model, we have adjusted the server pool setting in §4.2 and "optimized" the overflow trigger time parameters in §4.6 so that the ED-GW patient's waiting time performance can better match the empirical estimates. Such a parameter optimization is necessary because several model parameters are not able to be estimated from the NUH data directly (see more discussions in §§4.2 and 4.6). Admittedly, using such an "optimized" baseline model, one does not expect to predict well other performance measures such as the overflow proportion. As reported in the preceding paragraph, the simulation estimates of overflow proportions are actually within the ballpark of the empirical estimates, an indication that our model captures the inpatient flow operations at reasonably high fidelity.

There are certain performance measures that we choose not to calibrate, including the waiting time statistics for ICU-GW and SDA patients. For this, we offer two justifications. First, these performance measures are not the focus of this paper; see previous discussion in §2.3.1. Second, our sensitivity analysis shows that whether we can accurately replicate their waiting times has little impact on the waiting time statistics of ED-GW patients; the latter are the focus of this paper.

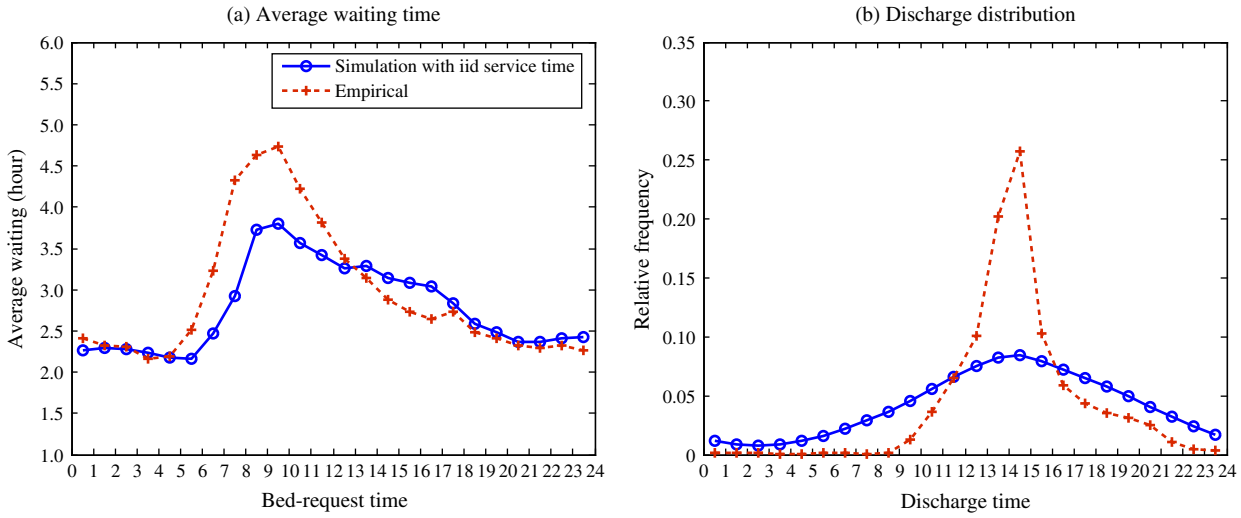
5.2. Models Missing Any of the Critical Features

To show the necessity of modeling the three critical features discussed in §3 (i.e., the two-time-scale service times, overflow mechanism, and allocation delays), we simulate three versions of the model, each missing one of the critical features. All other input settings for the three versions remain the same as we simulate the baseline scenario unless otherwise specified. Again, we compare the simulation estimates against the empirical performance in period 1.

5.2.1. Model With Conventional iid Service Times. The two-time-scale service time model proposed in §3.2 is contrary to the exogenous, iid service time model often used in the queueing literature. We compare an iid service time model with our proposed non-iid model. The iid model assumes the service time S to be the sum of two independent random variables: an integer variable corresponding to the floor of service time $\lfloor S \rfloor$, and a residual variable corresponding to $(S - \lfloor S \rfloor)$. For patients from the same class, we assume their integer parts and residual parts each form an iid sequence based on the empirical evidence. Since the two sequences are independent, the service times are iid. Although this iid exogenous service time model can reproduce service time distributions such as the one in Figure 5(a), it cannot reproduce the discharge distribution. As a result, the hourly waiting time statistics cannot be reproduced either. See the simulation output in Figure 13 for an illustration. Therefore, we believe that our new two-time-scale service time model is an important feature to capture inpatient flow operations.

5.2.2. Model Without Allocation Delays. Figure 14 compares the simulation and empirical estimates of the hourly average waiting time and hourly average queue length for ED-GW patients. In the simulation setting, no allocation delays are modeled. We can see that the hourly performance curves from simulation are completely different from the empirical estimates. In particular, note that the solid curve in Figure 14(b), which shows a rapid drop in the simulated average queue length between 11 A.M. and

Figure 13 **(Color online) Simulation Output from Using an iid Service Time Model**



3 P.M., contrasts sharply with the empirical (dashed) curve, which drops slowly after 2 P.M. The main reason for the rapid drop in the solid curve is that in period 1, between 11 A.M. and 3 P.M., the discharge rate increases in each hour until reaching the peak between 2 P.M. and 3 P.M. (see Figure 2(a)), and a waiting patient in the simulation is admitted into service immediately once a discharge occurs. Thus, Figure 14 suggests the existence of extra delays after bed discharges. In this simulation scenario, to make the daily average waiting time comparable to the estimate from the baseline scenario (2.82 hours), we decrease the numbers of servers listed in Table 1 while keeping all other settings the same as in the baseline scenario.

5.2.3. Model Without the Dynamic Overflow Policy. Section 3.3 has discussed the important role of overflow in achieving a QED regime for hospital operations. Furthermore, we find that adopting a dynamic overflow policy is also critical to replicate the empirical performance at NUH. Figure 15 compares empirical estimates of the hourly waiting time statistics with simulation estimates from a model with a static overflow trigger time $T = 4.0$ hours. Clearly, the model with this static overflow policy fails to capture the dynamics in NUH inpatient operations. In particular, note that under the static overflow policy, the simulation estimates of the average waiting time for patients arriving in the night (10 P.M. to 5 A.M. the next day) are approximately 4 hours, much higher than the empirical estimates. It is because in the simulation these night arrivals have to wait at least 4 hours for an overflow bed if no primary bed is available upon their arrivals, although a new primary bed is unlikely to become available within 4 hours due to the discharge pattern. We have tested other static

Figure 14 **(Color online) Simulation Output from a Model Without Allocation Delays**

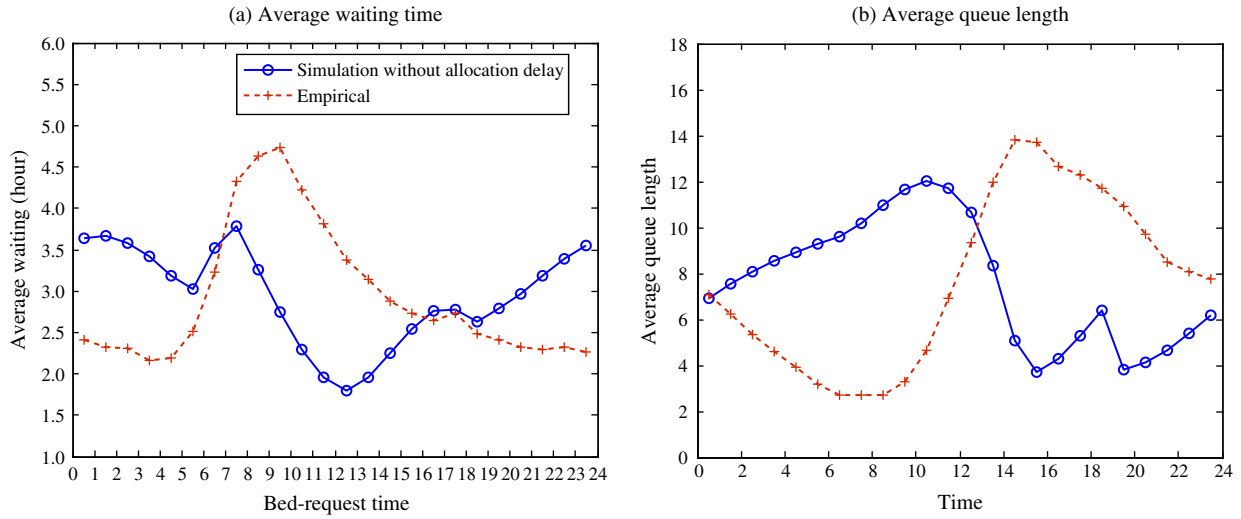
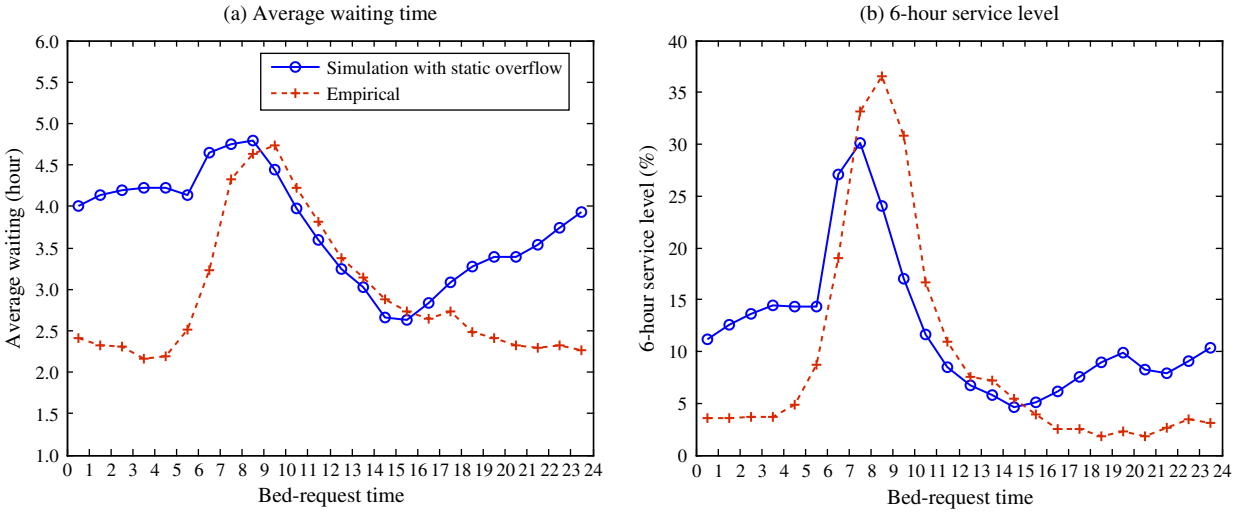


Figure 15 **(Color online) Simulation Output from Using a Static Overflow Policy with a Fixed Overflow Trigger Time $T = 4.0$ Hours**



overflow policies with the values of T ranging from 0 hours to 12 hours, and they cannot reproduce the empirical performance either.

6. Factors That Impact ED-GW Patients' Waiting Times

We do "what-if" analyses in this section and address the two research questions raised in the introduction, i.e., (i) quantify the impact of NUH's period 2 early discharge policy and (ii) identify operational policies that can reduce the waiting times of ED-GW patients. We focus on the impact of the tested policies on both the daily and hourly waiting time performance. In §6.1, we show that early discharge in period 2 has little impact on the daily and hourly waiting time statistics. In §6.2, we show that a hypothetical period 3 policy can flatten the hourly waiting time performance, but it has limited impact on reducing the daily waiting time statistics. In §6.3, we study policies that mainly impact the daily waiting time performance, such as increasing bed capacity and reducing LOS. In §6.4, we show that most of our gained insights are robust under sensitivity analysis. Finally, we explain in §6.5 why these policies have different impacts on the daily and hourly waiting time performance.

6.1. Period 2 Discharge Has a Limited Impact on Reducing Waiting Time Statistics

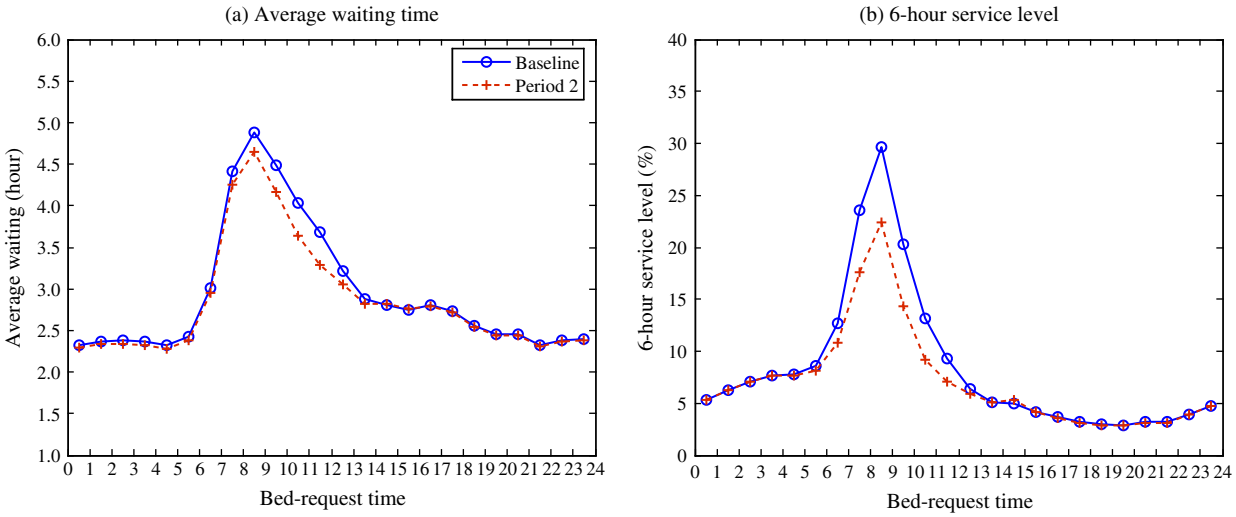
To evaluate the impact of NUH's period 2 early discharge policy, we simulate a scenario with the same inputs as in the baseline scenario, but using the discharge distribution estimated from period 2 data (i.e., using the dashed curve in Figure 2(a) instead of the solid curve). Figure 16 compares the simulation estimates of hourly waiting time statistics with those from the baseline scenario. In Figure 16(a), the hourly average waiting times show little difference between the two scenarios. In Figure 16(b), the hourly 6-hour service level exhibits some reduction for bed requests between 7 A.M. and 11 A.M.; e.g, the peak value is now 22% compared to 30% in the baseline scenario, but the values for other hours are almost identical in both scenarios. Not surprisingly, other performance measures from these two scenarios are almost identical. The daily average waiting time under this early discharge scenario is 2.75 hours, a 4-minute reduction, versus 2.82 hours in the baseline scenario. The 6-hour service level is 5.64% versus 6.29% in the baseline scenario. The overflow proportion is 21.88%, not significantly different from the baseline value of 21.70%.

To summarize, our model predicts that the period 2 early discharge policy has limited impact on reducing daily waiting time statistics and overflow proportions at NUH, and that this policy alone cannot flatten the waiting time performance throughout the day although it helps to reduce the peak hourly 6-hour service level. This prediction is consistent with our empirical observations of performance in periods 1 and 2; e.g., see Figure 1 in the introduction.

6.2. A Hypothetical Period 3 Policy Can Have a Significant Impact on Flattening Waiting Time Statistics

We consider a hypothetical discharge distribution that still discharges 26% of patients before noon as in period 2 but shifts the first discharge peak time between 8 A.M. and 9 A.M., i.e., three hours earlier than the first discharge peak time in period 2. Figure 2(b) plots this hypothetical discharge distribution. In addition, we assume a hypothetical allocation delay model: each allocation delay (pre- or post-allocation delay) follows a log-normal distribution with a constant mean, which is estimated from

Figure 16 **(Color online) Comparing Hourly Waiting Time Statistics Under the Baseline Scenario and Scenario with Period 2 Discharge Distribution**

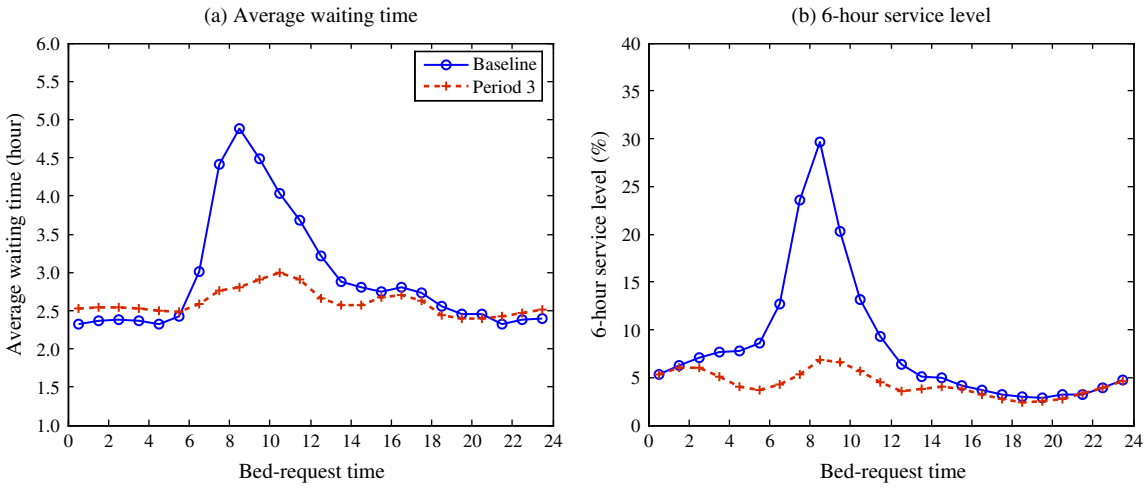


the empirical daily average. The estimated means of the pre- and post-allocation delays are 1.07 and 1.20 hours, respectively. We keep the same values of CV as in the baseline scenario; i.e., CV = 1 and CV = 0.6 for the pre- and post-allocation delays, respectively. We call the combination of the hypothetical discharge distribution and the hypothetical allocation delay model the "period 3 policy." The period 3 policy has not been implemented yet at NUH and may not be fully practical. We call it the period 3 policy because it has the potential to be implemented in the future. For consistency, we call the combination of the period 2 discharge distribution (which has been implemented) and the time-varying allocation delay model (see §4.7) the "period 2 policy."

Figure 17 compares the hourly waiting time statistics between the baseline scenario and the hypothetical period 3 scenario. Under the period 3 policy, patients requesting beds in the morning (7 A.M. to noon) experience similar average waiting times (2.76–2.99 hours) as the daily average (2.59 hours), but the daily average is only 13 minutes lower than the daily average in the baseline scenario. The peak value of the hourly 6-hour service level drops from 30% under the baseline scenario to 6.9% under the period 3 policy, with the daily 6-hour service level dropping from 6.29% to 4.02%. The overflow proportion drops slightly, from 21.70% under the baseline scenario to 21.40% under the period 3 policy.

Compared to the period 2 policy, the period 3 policy requires achieving both a more aggressive early discharge distribution and allocation delays that are time stable with constant means throughout the day. Simulation results show that when either component is missing (i.e., only implementing the aggressive early discharge policy or only stabilizing the allocation delays), the average waiting times for morning bed requests are still approximately 1–2 hours longer

Figure 17 **(Color online) Comparing Hourly Waiting Time Statistics Under the Baseline Scenario and Scenario with the Period 3 Policy**



Note. This hypothetical discharge distribution has a first peak at 8–9 A.M. and constant-mean allocation delays.

than the daily average, and the waiting time performance is not approximately flattened.

In summary, our model predicts that this hypothetical period 3 policy can eliminate the excessively long waiting times for ED-GW patients requesting beds in the morning. Simultaneous implementation of both the aggressive early discharge policy and allocation delay stabilization is necessary for the period 3 policy to achieve an approximately time-stable performance in waiting times. This suggests that coordinating other resources (such as ward nurses) to stabilize allocation delays is equally important when a hospital implements the early discharge policy to eliminate the long wait in the morning.

6.2.1. Findings from Other Early Discharge Scenarios. To obtain more insights into the impact of discharge timing, we test other hypothetical discharge distributions combined with the time-varying or constant-mean allocation delay models. Section 1 of the online supplement details the tested policies and simulation results. Here, we highlight two main observations that are consistent with what we see from the period 3 policy.

First, an early discharge policy mainly impacts the time-of-day pattern of the waiting time performance. Several tested combinations of early discharge distributions and constant-mean allocation delays can flatten the waiting time performance, but neither early discharge alone nor stabilizing allocation delays alone can. Moreover, we find that the timing of the first discharge peak has a major impact on flattening the performance. For example, if the hospital retains the first discharge peak time between 11 A.M. and noon as in period 2, even pushing 75% of patients to discharge before noon and stabilizing the allocation delays cannot approximately flatten the waiting time performance.

Second, an early discharge policy has limited impact on the daily average waiting time and overflow proportion in the NUH setting. In particular, we test a discharge distribution with every patient discharged as early as midnight to study the largest improvement that an early discharge policy might bring. When the mean allocation delays are constant, the daily average waiting time under this extreme early discharge scenario is 2.42 hours (a 24-minute reduction from the baseline scenario), and the overflow proportion is 19.73% (only a 2 percentage point reduction from the baseline scenario).

The reason that early discharge policies mainly impact the hourly performance but not the daily performance will be explained in §6.5.

6.3. Policies Impact the Daily Waiting Time Statistics and Overflow Proportion

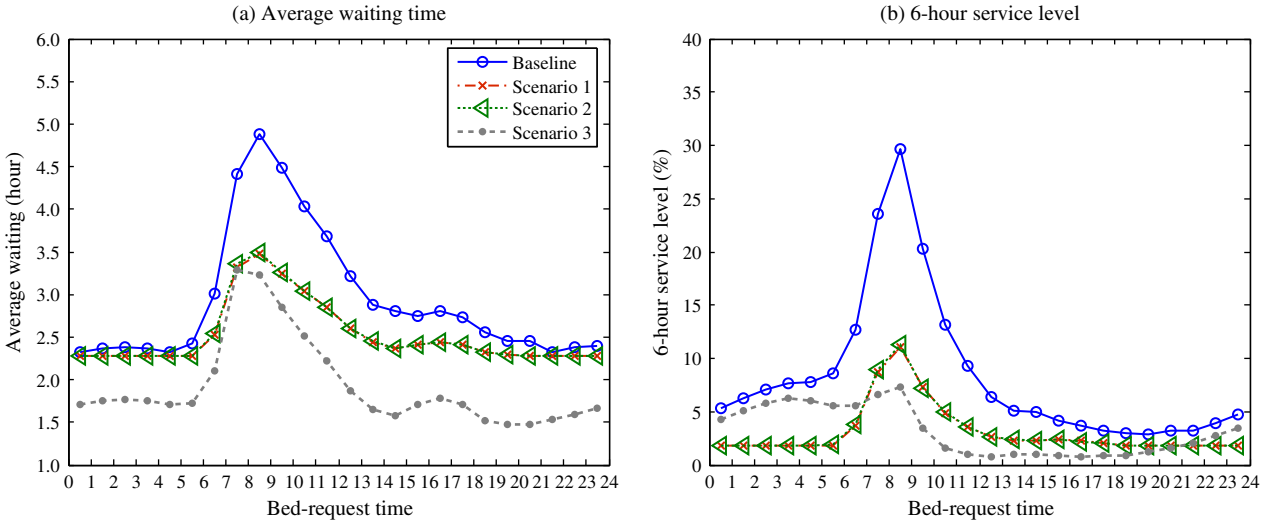
In this section, we show three policies that can significantly reduce the daily waiting time statistics and overflow proportions. They are increasing the bed capacity, reducing the LOS, and reducing the mean allocation delays. Specifically, we consider three scenarios. In the first one, we increase the number of servers from 563 (baseline) to 632 so that the utilization rate is reduced from 89.2% to 79.4% (a 10 percentage point reduction). In the second one, we eliminate excessively long LOS by limiting each patient to stay in the hospital for a maximum of 14 days. The utilization rate is thereby reduced to 78.5%, close to that in the first scenario. In the third scenario, we reduce the mean pre- and post-allocation delays by 30 minutes each. In each scenario, we use the baseline (period 1) discharge distribution and assume the constant-mean allocation delay model; all other settings not specified here remain the same as in the baseline scenario.

The daily average waiting times are 2.45, 2.46, and 1.80 hours in the first, second, and third scenario, respectively. The daily 6-hour service levels are 2.60%, 2.60%, and 2.31%, respectively; the overflow proportions are 12.04%, 12.05%, and 21.30%, respectively. Figure 18 plots the hourly waiting time statistics under the three scenarios. Compared to the baseline scenario, a 10% capacity increase results in a significant reduction in the overflow proportion (a 9.7 percentage point reduction) but only reduces the daily average waiting time by 22 minutes. Reducing the LOS shows a similar impact since it is essentially equivalent to creating more capacity. A total reduction of one hour in the mean pre- and post-allocation delays leads to a reduction of one hour in the daily average waiting time, although it has limited impact on reducing the overflow proportion. In all three scenarios, the daily 6-hour service levels are significantly reduced.

Furthermore, we see from Figure 18 that, in all three scenarios, the hourly average waiting time is not stabilized; i.e., the average waiting time for patients requesting beds between 7 A.M. and 11 A.M. is still approximately 1–2 hours longer than the daily average. The hourly 6-hour service level, however, appears to be more time stable than the average waiting time for each scenario, especially considering that the peak value is 30% of the baseline. Note that if we increase the bed capacity to 707 beds (the utilization rate reduces to 71.0%), the waiting time curves can be approximately stabilized, whereas reducing the mean allocation delays to 0 still cannot achieve a time-stable performance.

In summary, our model predicts that reducing the mean allocation delays can significantly reduce the daily average waiting times, whereas increasing the bed capacity or reducing the LOS mainly impacts the overflow proportion in the NUH setting. Moreover, these policies have less impact on the time-of-day pattern of waiting time performance, nor do

Figure 18 **(Color online) Hourly Waiting Time Statistics Under Three Scenarios**



Notes. Scenario 1: increasing bed capacity by 10%; scenario 2: assuming each patient can stay in the hospital for a maximum of 14 days; scenario 3: reducing the mean pre- and post-allocation delays by 30 minutes each. In each scenario, we use the period 1 discharge distribution and assume the constant-mean allocation delay model.

they necessarily flatten the hourly waiting time performance. Our results also suggest that at NUH the 2–3 hours average waiting time mainly comes from secondary bottlenecks other than bed unavailability, such as inadequate nurse staffing. This finding is consistent with the observation in a recent paper (Pines et al. 2011a) that the long waiting time of ED-GW patients may not be caused by a lack of inpatient beds but rather by other inefficiencies that slow the transitions of care between different hospital units. In the following section, we will evaluate the impact of increasing capacity and reducing LOS in a more capacity-constrained setting.

6.4. Sensitivity Analysis

To examine the robustness of the insights we have gained so far, we test five policies—the period 2 and period 3 policies and the three alternative policies described in the previous section—under different model settings for sensitivity analysis. These settings include using alternative arrival models; changing the priority among ICU-GW, SDA, and ED-GW patients; choosing different values for the normal allocation probability $p(t)$′ and using alternative server pool settings and overflow policies. The simulation results show that the insights we gained are robust under the tested model variations.

In addition, to evaluate the five policies when the system load is high, we increase the daily arrival rates of ED-GW patients by 7%, similar to the increase we empirically observed from period 1 to period 2. When all other settings are kept the same as in the baseline, utilization under the increased arrival rate assumption becomes 93%, and the daily average waiting time and 6-hour service level become 4.37 hours and 18.60%, respectively. We test the five policies under the increased arrival rate assumption. We find that the insights we have gained are still robust in this capacity-constrained setting, except that (a) increasing capacity by 10% or reducing the LOS now shows a significant reduction in daily average waiting times (from 4.37 to approximately 2.5 hours); and (b) the period 3 policy can also greatly reduce the daily waiting time statistics because of its side effect of reducing the LOS, which results from using different LOS distributions between ED-GW patients admitted in the A.M. and P.M. (see Table 3). Sections 2, 3, and 5 of the online supplement detail all the experimental setting and simulation results of the sensitivity analysis discussed in this section.

6.5. Intuition About the Insights Gained

Our evaluated policies show different impacts on the daily and hourly waiting time performance. The reason lies in the separation of time scales, which is captured by our two-time-scale service time model in Equation (1). We now provide some intuition to explain the findings we have obtained so far. A more mathematical explanation can be obtained through the analytical framework developed in Dai and Shi (2014).

There are two types of wait in our model. (i) The total number of discharges in one day is less than the total number of arrivals in that day, and therefore some patients have to wait until the next day to get a bed. (ii) Within a day, the discharge timing is so late that morning arrivals have to wait several hours (i.e., until the afternoon) to get a bed. The first type of

wait, reflected in the daily waiting time performance, can be affected by the daily arrival rate, LOS distributions, and bed capacity. The second type of wait, reflected in the time-of-day (hourly) waiting times, can be affected by the time-varying arrival rates and discharge patterns.

Clearly, merely shifting the discharge timing earlier can eliminate or reduce the second type of wait, but not the first. Thus, early discharge policies can flatten hourly waiting time curves, but they have limited impact on further reduction of daily waiting times (when there is no side effect in reducing the LOS). Moreover, to achieve the flattening effect, the early discharge policy needs to ensure that a modest number of patients be discharged early enough (before 9–10 A.M. in the NUH setting as suggested by our simulation results) that beds become available before the queue starts to build up in the morning. This also explains why period 2 policy has a limited impact on flattening the hourly waiting time curves, since its first discharge peak starts at 11 A.M., which is not early enough.

In comparison, increasing capacity or reducing LOS helps to reduce the first type of wait and thus can reduce daily waiting time statistics. This effect is particularly significant in a capacity-constrained setting. Even when bed utilization is low but not excessively low (not lower than 71% in the NUH setting), many morning arrivals can still experience the second type of wait because not enough patients are discharged early in the morning. This is why increasing capacity does not necessarily flatten the hourly waiting time curves.

Furthermore, the insights we gained here can help hospital managers to make better decisions when choosing among different policies. For example, when the system load is very high and a significant proportion of the patients have to wait overnight, merely shifting the discharge timing may not help much, since the long wait mainly comes from insufficient capacity; increasing capacity (or reducing LOS) would bring more significant impact. When the capacity issue is solved and the system is moderately loaded, then managers can implement the early discharge policy to further improve operations and get rid of the peak wait in the morning.

7. Concluding Remarks and Future Research

We have proposed a high-fidelity stochastic network model for inpatient flow management that can be used as a tool to quantify the impact of various operational policies. In particular, the model captures time-of-day waiting time performance for ED-GW patients and enables us to identify policies that can reduce or flatten waiting times. Our model predicts that a hypothetical period 3 policy (and similar policies with certain early discharge distributions and constant-mean allocation delays) can achieve time-stable waiting time statistics throughout the day but has limited impact on the daily average waiting time and overflow proportion in the NUH setting. Our model also predicts that reducing the mean allocation delays, increasing bed capacity, or reducing LOS mainly affects the daily average waiting time and overflow proportion; however, these three policies have less impact on the time-of-day pattern of waiting time performance, and they do not necessarily stabilize waiting time statistics. These insights can help hospital managers to choose among different policies to implement, depending on the choice of objective, such as to reduce the peak wait in the morning or to reduce daily waiting time statistics.

Readers should be aware of two issues when interpreting these findings. First, we focus on evaluating the impact of discharge policies and other policies on the waiting time performance of ED-GW patients in this paper. There could be other benefits of these policies that our paper has not modeled. For example, it is believed that early discharge allows more flexibility to transfer patients from the ICU to GWs when ICU wards become congested. Second, regarding the impact on the waiting times, the evaluation of these policies is based on predictions from the populated NUH model and comparison to the baseline scenario. Thus, our findings may not be always generalizable to other hospital settings. Section 1.4 of the online supplement shows an example where the period 2 policy can have more significant benefits in a different setting.

On the implementation level, we recognize the challenges in implementing the period 3 policy in practice. On the one hand, discharging patients as early as 8–9 A.M. is difficult since physicians and nurses are busy with the morning rounds at about this time. On the other hand, shifting the discharge timing can impact the availability of other resources (e.g., nurses, cleaning crews), and managing these secondary resources to stabilize allocation delays is equally important. This would require coordination throughout the entire hospital and proper staffing at different units at various hours of the day. Although the period 3 policy is purely hypothetical and may not be completely practical, we believe it can serve as a goal for hospital managers to aim at if they intend to eliminate excessively long waiting times for morning bed requests.

More importantly, our model provides an efficient tool to evaluate the impact of a spectrum of policies that lie between period 2 and period 3 policies. Based on the outcomes and costs of implementation,

hospital managers can choose the desired levels of effort to implement these policies. Our model, particularly under the current allocation delay setting, also provides a possible framework to manage the complicated interaction among bed and other secondary resources. It allows one to understand the impact of changing one factor (e.g., discharge timing) while keeping all other factors fixed. Finally, besides evaluating discharge policies, our model has the potential to be developed into a decision-support system that allows hospital managers to quantify and evaluate the trade-off between the benefit of reducing ED overcrowding and the cost of implementing a number of operational and strategic policies.

7.1. Future Work

The proposed model on inpatient flow management in this paper can be used for other studies that intend to integrate the ED and inpatient department operations together. Our model can potentially be extended in several directions.

First, detailed studies are needed to further understand the secondary bottlenecks during the transfer process from the ED to the wards and their interaction with bed availability. Then one can explicitly incorporate these bottlenecks into the model to identify strategies on staffing and scheduling that can reduce allocation delays. The two-queue model proposed in Yankovic and Green (2011) appears relevant to this line of research.

Second, our proposed model is a parallel-server system with a single-pass routing structure. In particular, we do not model ICU-type wards and patient flows within ICU-type wards in our system, because the data requirements to model them would be at another level and are beyond the scope of this paper. An extension would be to build upon this paper and recent studies on ICU management (Chan et al. 2014, Kim et al. 2015) to model both ICU-type wards and general wards as a stochastic network that has internal routings between these wards. The extended model could study waiting times for ED-GW and ICU-GW patients, as well as waiting times for ED-ICU patients or GW-ICU patients, and could better evaluate the impact of early discharge and other policies on patients besides ED-GW patients.

Third, considering day-of-week phenomena is another important extension to make the model more realistic. Currently, we assume that ED-GW patients have a stable daily arrival volume without differentiating days of the week. We also assume that elective admissions are stationary by day, although recent work pointed out that elective schedule is actually the main source of daily occupancy variation in many hospitals (Helm and Van Oyen 2014). Our model can be extended to predict day-of-week performance and help to design a better elective schedule.

Fourth, our model assumes a simple dynamic overflow policy, although the actual overflow decisions are much more complicated. Besides, the model does not capture many constraints during the real bed assignment process such as patient gender. Better modeling the current overflow practice and proposing guidelines to improve bed assignment decisions will be left for future study.

Finally, obtaining structural insights into the impact of many policies, such as discharge timing and overflow trigger time, by simulations alone is difficult. There is a need to develop analytical methodology, not purely simulations, to predict performance measures that depend on hour of day. We believe that the model proposed in this paper will stimulate new analytical research to develop tools to study a new class of models. See, for example, some preliminary tools developed in Dai and Shi (2014).

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/mnsc.2014.2112.

Acknowledgments

The authors give special thanks to James Ang at National University of Singapore for bringing the groups together for this collaboration and all the help on an earlier draft, and to Xin Jin and Wei Ping Goh at National University Hospital for many stimulating discussions and enthusiastic support in every aspect of our research. The authors also thank Assaf Zeevi, the associate editor, and three anonymous referees for their valuable comments and suggestions, which have greatly improved the exposition of this paper. This research was supported in part by the National Science Foundation [Grants CMMI-1030589, CNS-1248117, and CMMI-1335724], the National University of Singapore Academic Research Fund [Grant R-314-000-082-112], the National University of Singapore Medicine-Business Seed Grant [Grant C-311-004-003-001], the Ministry of Health [Grant HSRG/0002/2010], and the National Natural Science Foundation of China [Grant 71102082].

References

Allon G, Deo S, Lin W (2013) The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Oper. Res.* 61(3):544–562.

Armony M, Israelit S, Mandelbaum A, Marmor Y, Tseytlin Y, Yom-Tov G (2015) On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems.* Forthcoming.

Bair AE, Song WT, Chen Y-C, Morris BA (2010) The impact of inpatient boarding on ED efficiency: A discrete-event simulation study. *J. Medical Systems* 34(5):919–929.

Bertsekas D, Gallager R (1992) *Data Networks* (Prentice-Hall, Englewood Cliffs, NJ).

Birjandi A, Bragg LM (2008) *Discharge Planning Handbook for Healthcare: Top 10 Secrets to Unlocking a New Revenue Pipeline* (Productivity Press, New York).

Borghans I, Heijink R, Kool T, Lagoe RJ, Westert GP (2008) Benchmarking and reducing length of stay in Dutch hospitals. *BMC Health Services Res.* 8(1), doi: 10.1186/1472-6963-8-220.

Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L (2005) Statistical analysis of a telephone call center. *J. Amer. Statist. Assoc.* 100(469):36–50.

Centers for Disease Control and Prevention (2011) Health, United States, 2010: With special feature on death and dying. Report, Centers for Disease Control and Prevention, Atlanta. http://www.cdc.gov/nchs/data/hus/hus10.pdf.

Chan C, Yom-Tov G, Escobar GJ (2014) When to use speedup: An examination of service systems with returns. *Oper. Res.* 62(2): 462–482.

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

Dai JG, Lin W (2005) Maximum pressure policies in stochastic processing networks. *Oper. Res.* 53(2):197–218.

Dai JG, Shi P (2014) A two-time-scale approach to time-varying queues for hospital inpatient flow management. Working paper, Cornell University, Ithaca, NY. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2489533.

de Bruin AM, van Rossum AC, Visser MC, Koole GM (2007) Modeling the emergency cardiac in-patient flow: An application of queuing theory. *Health Care Management Sci.* 10:125–137.

Department of Health, United Kingdom (2004) Achieving timely simple discharge from hospital: A toolkit for the multi-disciplinary team. Report, Department of Health, London. http://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_4088367.pdf.

Feldman Z, Mandelbaum A, Massey WA, Whitt W (2008) Staffing of time-varying queues to achieve time-stable performance. *Management Sci.* 54(2):324–338.

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* 5(2):79–141.

Green L (2002) How many hospital beds? *Inquiry* 39(4):400–412.

Green L (2006) Queueing analysis in healthcare. Hall RW, ed. *Patient Flow: Reducing Delay in Healthcare Delivery*, International Series in Operations Research and Management Science, Vol. 91 (Springer, New York), 281–307.

Griffin J, Xia S, Peng S, Keskinocak P (2011) Improving patient flow in an obstetric unit. *Health Care Management Sci.* 15(1):1–14.

Hall MJ, DeFrances CJ, Williams SN, Golosinskiy A, Schwartzman A (2010) National hospital discharge survey: 2007 summary. *Natl. Health Stat. Report* 29:1–20.

Hall RW (2012) Bed assignment and bed management. Hall R, ed. *Handbook of Healthcare System Scheduling*, International Series in Operations Research and Management Science, Vol. 168 (Springer, New York), 177–200.

Hall RW, Belson D, Murali P, Dessouky M (2006) Modeling patient flows through the healthcare system. Hall RW, ed. *Patient Flow: Reducing Delay in Healthcare Delivery* (Springer, New York), 1–44.

Harrison JM (2003) Brownian models of open processing networks: Canonical representation of workload. *Ann. Appl. Probab.* 13(1): 390–393.

Helm J, Van Oyen M (2014) Design and optimization methods for elective hospital admissions. *Oper. Res.* 62(6):1265–1282.

Helm JE, AhmadBeygi S, Van Oyen MP (2011) Design and analysis of hospital admission control for operational effectiveness. *Production Oper. Management* 20(3):359–374.

Hoot NR, Aronsky D (2008) Systematic review of emergency department crowding: Causes, effects, and solutions. *Ann. Emergency Medicine* 52(2):126–136.

Howell E, Bessman E, Kravet S, Kolodner K, Marshall R, Wright S (2008) Active bed management by hospitalists and emergency department throughput. *Ann. Internal Medicine* 149(11):804–810.

Jacobson SH, Hall SN, Swisher JR (2006) Discrete-event simulation of health care systems. Hall RW, ed. *Patient Flow: Reducing Delay in Healthcare Delivery*, International Series in Operations Research and Management Science, Vol. 91 (Springer, New York), 211–252.

Khanna S, Boyle J, Good N, Lind J (2011) Impact of admission and discharge peak times on hospital overcrowding. Hansen DP, Maeder AJ, Schaper LK, eds. *Health Informatics: The Transformative Power of Innovation*, Studies in Health Technology and Informatics, Vol. 168 (IOS, Amsterdam), 82–88.

Kim S-H, Chan CW, Olivares M, Escobar G (2015) ICU admission control: An empirical study of capacity allocation and its implication on patient outcomes. *Management Sci.* 61(1):19–38.

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

Kumar PR (1993) Re-entrant lines. *Queueing Systems* 13(1–3):87–110.

Law AM, Kelton DW (2000) *Simulation Modelling and Analysis* (McGraw-Hill Education, Columbus, OH).

Lewis PA, Shedler GS (1978) Simulation methods for Poisson processes in nonstationary systems. *Proc. 10th Conf. Winter Simulation*, WSC '78, Vol. 1 (IEEE Press, Piscataway, NJ), 155–163.

Litvak E, Long MC, Cooper AB, McManus ML (2001) Emergency department diversion: Causes and solutions. *Acad. Emergency Medicine* 8(11):1108–1110.

Liu Y, Whitt W (2012) Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.* 60(6): 1551–1564.

Liu SW, Thomas SH, Gordon JA, Hamedani AG, Weissman JS (2009) A pilot study examining undesirable events among emergency department-boarded patients awaiting inpatient beds. *Ann. Emergency Medicine* 54(3):381–385.

Mandelbaum A, Momcilovic P, Tseytlin Y (2012) On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers. *Management Sci.* 58(7):1273–1291.

National University Hospital (2011) BMU Training Guide: Inpatient Operations, National University Hospital, Singapore.

Pines JM, Batt RJ, Hilton JA, Terwiesch C (2011a) The financial consequences of lost demand and reducing boarding in hospital emergency departments. *Ann. Emergency Medicine* 58(4): 331–340.

Pines JM, Iyer S, Disbot M, Hollander JE, Shofer FS, Datner EM (2008) The effect of emergency department crowding on patient satisfaction for admitted patients. *Acad. Emergency Medicine* 15(9):825–831.

Pines JM, Hilton JA, Weber EJ, Alkemade AJ, Al Shabanah H, Anderson PD, Bernhard M, et al. (2011b) International perspectives on emergency department crowding. *Acad. Emergency Medicine* 18(12):1358–1370.

Powell ES, Khare RK, Venkatesh AK, Van Roo BD, Adams JG, Reinhardt G (2011) The relationship between inpatient discharge timing and emergency department boarding. *J. Emergency Medicine* 42(2):186–196.

Ramakrishnan M, Sier D, Taylor P (2005) A two-time-scale model for hospital patient flow. *IMA J. Management Math.* 16(3):197–215.

Schneider S, Zwemer F, Doniger A, Dick R, Czapranski T, Davis E (2001) Rochester, New York: A decade of emergency department overcrowding. *Acad. Emergency Medicine* 8(11):1044–1050.

Shi P (2013) Stochastic modeling and decision making in two healthcare applications: Inpatient flow management and influenza pandemics. Ph.D. dissertation, Georgia Institute of Technology, Atlanta.

Shi P, Dai JG, Ding D, Ang J, Chou MC, Jin X, Sim J (2014) Patient flow from emergency department to inpatient wards: Empirical observations from a Singaporean hospital. Working paper, Purdue University, West Lafayette, IN. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2517050.

Sinagpore Ministry of Health (2012) Waiting time for admission to ward. Accessed October 30, 2014, https://www.moh.gov.sg/content/moh_web/home/statistics/healthcare_institutionstatistics/Waiting_Time_for_Admission_to_Ward.html.

Singer AJ, Thode J, Henry C, Viccellio P, Pines JM (2011) The association between length of emergency department boarding and mortality. *Acad. Emergency Medicine* 18(12):1324–1329.

Teow K, El-Darzi E, Foo C, Jin X, Sim J (2012) Intelligent analysis of acute bed overflow in a tertiary hospital in Singapore. *J. Medical Systems* 36(3):1873–1882.

Thompson S, Nunez M, Garfinkel R, Dean M (2009) Efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges. *Oper. Res.* 57(2):261–273.

United States General Accounting Office (2003) Hospital emergency departments: Crowded conditions vary among hospitals and communities Report, United States General Accounting Office, Washington, DC.

Vericourt Fd, Jennings OB (2011) Nurse staffing in medical units: A queueing perspective. *Oper. Res.* 59(6):1320–1331.

Wong HJ, Morra D, Caesar M, Carter MW, Abrams H (2010) Understanding hospital and emergency department congestion: An examination of inpatient admission trends and bed resources. *Canadian J. Emergency Medicine* 34(1):18–26.

Yancer DA, Foshee D, Cole H, Beauchamp R, de la Pena W, Keefe T, Smith W, Zimmerman K, Lavine M, Toops B (2006) Managing capacity to reduce emergency department overcrowding and ambulance diversions. *Joint Commission J. Quality Patient Safety* 32(5):239–245.

Yankovic N, Green LV (2011) Identifying good nursing levels: A queuing approach. *Oper. Res.* 59(4):942–955.

Yao DD (1994) *Stochastic Modeling and Analysis of Manufacturing Systems*, Springer Series in Operations Research (Springer, New York).

Zacharias C, Armony M (2015) Joint panel sizing and appointment scheduling in outpatient care. Working paper, University of Miami, Coral Gables, FL.

Zeltyn S, Marmor YN, Mandelbaum A, Carmeli B, Greenshpan O, Mesika Y, Wasserkrug S, et al., eds. (2011) Simulation-based models of emergency departments: Operational, tactical, and strategic staffing. *ACM Trans. Model. Comput. Simul.* 21(4):Article 24.