Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0000-0000 | EISSN 0000-0000 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000


# Models and Insights for Hospital Inpatient Operations: Time-of-Day Congestion for ED Patients Awaiting Beds *

Pengyi Shi
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, pengyishi@gatech.edu

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

J. G. Dai
School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853; on leave from H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, jd694@cornell.edu

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

Joe Sim
NUS Yong Loo Lin School of Medicine and NUS Business School, National University of Singapore, and National University Hospital, 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 ED before being admitted. Based on an empirical study at a Singaporean hospital in the Companion Paper [48], 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 her admission and discharge times and on her 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 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 non-primary 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 *Period 3* policy can eliminate excessive waiting for those patients who request beds in mornings. The policy incorporates the following components: a discharge distribution with the first discharge peak between 8 and 9am 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 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.

*Key words*: inpatient flow management, early discharge, time-dependent waiting time, stochastic network model, ED boarding

* Original Title: "Hospital Inpatient Operations: Mathematical Models and Managerial Insights"

# 1. Introduction

Inpatient beds are one of the most critical resources in hospitals. Inpatient flow and bed management has crucial impacts on hospital operations [22], especially on emergency department (ED) crowdedness [4, 28, 36, 46, 55]. Prolonged waiting time for admission to inpatient wards, also known as ED boarding, has been identified as a key contributor to ED overcrowding worldwide [27, 42, 53]. 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. Though the model is built upon an extensive empirical study at one Singaporean hospital, we believe the modeling framework can be adapted to other hospitals based on the similarity in many empirical observations between this hospital and others.

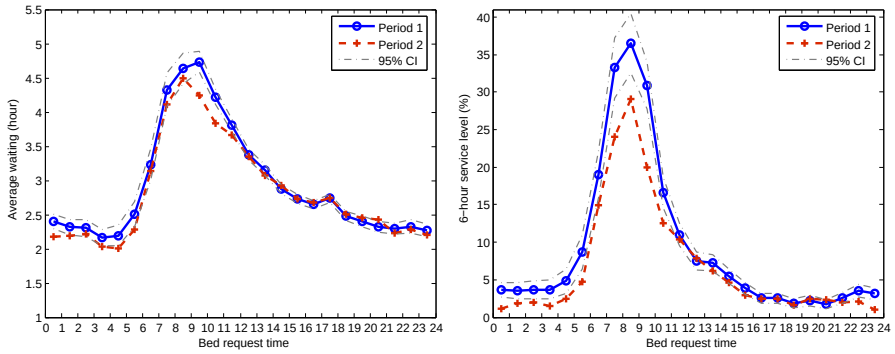
## 1.1. Motivation and research questions

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 about 1000 beds to serve patients admitted from ED and other sources. At NUH, around 20% of patients visiting 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 ward* of an ED-GW patient, or simply the *waiting time* in the rest of the paper, 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 is admitted to a GW.

**Time-of-day waiting time performance**

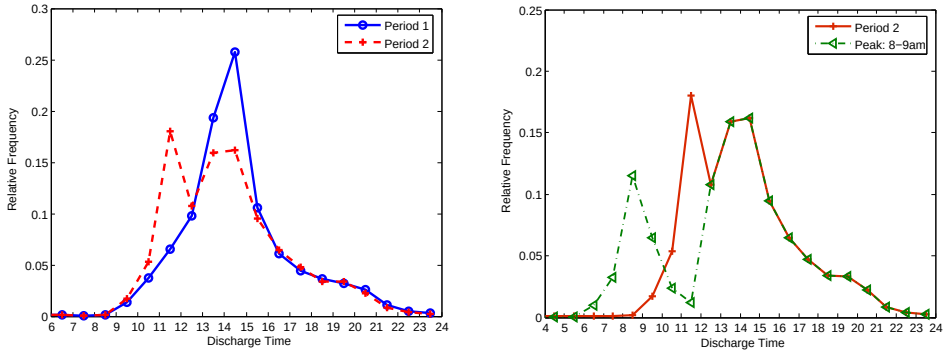
From January 1, 2008 to June 30, 2009, called Period 1 in this paper, the average waiting time at NUH is 2.82 hours (169 minutes), which does not seem to be very long. However, this level of complacency immediately evaporates if we examine the waiting times of patients requesting beds in mornings. The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

While no patient likes any waiting, 6 hours or more is extremely undesirable, not only because patients can get very frustrated during the long wait [43], but also because of the adverse outcome associated with it. Liu et al. [37] and Singer et al. [50] have discovered that patients who waited longer than 6 hours after their admission decisions have been made are more likely to experience longer inpatient stay, higher mortality rates, and other undesirable events in 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 ED, which lead to ED overcrowding and sometimes ambulance diversion [1]. Thus, it is important for hospitals to eliminate the excessive amount of waiting, especially for morning bed-requests.




(a) Average waiting times (b) 6-hour service level

**Figure 1** **Hourly waiting time statistics for ED-GW patients; 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 7am and 8am. The 95% confidence intervals are plotted for Period 1 curves.**




(a) Discharge distributions in Periods 1 and 2 (b) Period 2 discharge distribution and a hypothetical discharge distribution

**Figure 2** **Discharge time distributions in Periods 1 and 2, and a hypothetical discharge distribution with the first peak at 8-9am and 26% patients being discharged before noon. Each dot represents the fraction of patients who are discharged during that hour. The values in the first 4 hours are nearly zero in all three distributions and are not displayed.**

### 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 2a plots the discharge distribution of patients from general wards at NUH in Period 1. Clearly, the peak discharge hour is between 2pm and 3pm. Therefore, many admissions must wait until after 3pm, while bed-requests of ED-GW patients can occur during the entire day (see the solid curve in Figure 7 in Section 4.1). In other words, 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 at NUH. Similar waiting time curves have been observed in other hospitals (see Figure 30 of [2]), and so have the number of patients waiting at different time of a day [22, 44]. Meanwhile, the bed-request and discharge patterns

in many other hospitals are also similar to what we observed at NUH; see, e.g., Table 1 in [44] and Figure 6 in [2]. Studies in literature [6, 56] and government agencies [15] 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 2a displays the new discharge distribution. A morning discharge peak arises, occurring between 11am 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 1a and 1b 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% [48]. 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.

**Modeling.** For the first contribution, we develop a new stochastic network model which 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. In order for the model to be able to capture the inpatient operations at hourly resolution, we find 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 which 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 Section 3. Time-varying $M_t/GI/n$ queues or their network versions, where the arrival process is Poisson with time-varying arrival rate and the service times are iid, have been used in literature to model hospital operations; see, for example, [1, 19, 32]. Despite our best efforts, we are not able to reproduce the time-dependent performance curves using these models. See Section 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 [14], 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 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 *sub-groups* of patients, in addition to the aggregated impact on all patients. Moreover, as pointed out by Armony et al. [2], 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 a understanding of the time-varying dynamics of inpatient flow.

**Practice.** The second contribution is that, 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 8am and 9am, and still keep 26% discharge before noon (see the dash-dotted curve in Figure 2b) and (ii) meanwhile stabilize the time-varying allocation delays, then the hourly waiting time curves can be approximately flattened (see Figure 17). 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 and reducing the mean allocation delays; these policies do not necessarily flatten the hourly waiting time curves though. Finally, we provide some intuition to explain the different impacts on the

hourly and daily waiting time performance of these policies resulting from the separation of time scales, a phenomenon captured by our new service time model. See Section 6 for the details.

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. [44], 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 [14] 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 which depend on the *entire distribution* of the hourly patient count. Therefore, some of the managerial insights generated in [44] 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, [30] classifies admission data from 23 Australian hospitals into five categories based on the relative timing of daily admission and discharge curves, and uses 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, [2] and [23] conduct detailed studies of patient flow in various departments at an Israeli and a US hospital, respectively. Readers are also referred to the many articles cited in these two papers for further references. Armony et al. [2] 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 [18, 29, 59]. Compared to the rich literature on patient flow models of ED, inpatient flow management and the interface between ED and inpatient wards have received less attention; see the same discussion in Section 4 of [2]. Related works on inpatient operations include capacity allocation and flow improvement in specialized hospitals or wards [9, 12, 19, 20], ward nurse staffing [54, 57], bed assignment and overflow [39, 52], and elective admission control and design [25, 26]. Note that Yankovic and Green [57] 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 the allocation delays.

Stochastic network models have been a common tool to study manufacturing, communication and service systems [5, 17, 58]. 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. [16] and recent work by Liu and Whitt [38] propose staffing algorithms to achieve time-stable performance. Unlike call center models, our hospital model has extremely long service times with an average of about 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.

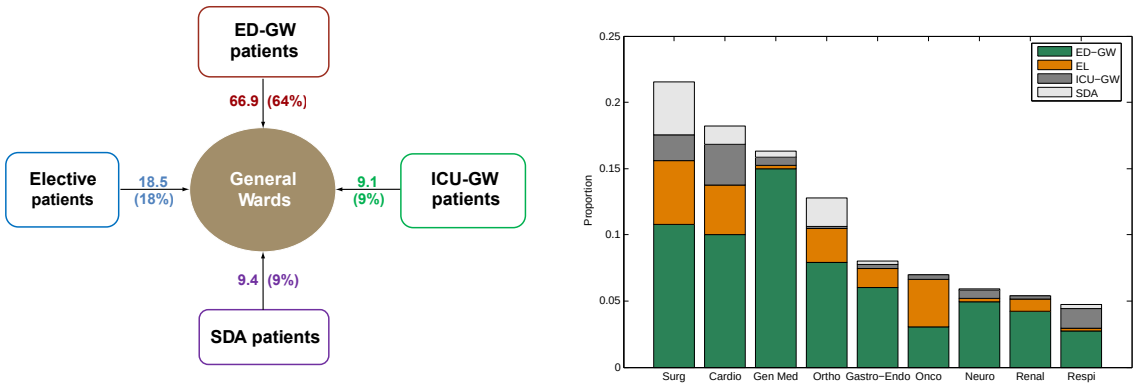
The remainder of this paper is organized as follows. In Section 2, we give a brief description of the NUH inpatient department and the performance measures we focus on. In Section 3, we introduce the general framework of our proposed stochastic network model that captures the inpatient flow operations. In Section 4, we populate the proposed stochastic network model with NUH data. In Section 5, we verify the populated model by comparing the model output with empirical performance. In Section 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 Section 7.

## 2. NUH inpatient department

This section briefly describes the operations of the NUH inpatient department. We focus on 19 *general wards* (GW's), which exclude 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 referred as *floor bed* in US hospitals. The total number of general beds at NUH ranges from 555 to 638 between January 1, 2008 and December 31, 2010. The precise definition of GW and reasons we exclude other wards from GW's are presented in the Companion Paper [48].

### 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 treatments in the ED and need to be admitted into a general ward. ICU-GW patients are those patients who are initially admitted to ICU-type wards (from either ED or other external sources) and are later transferred to general wards. 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 into 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 surgery, SDA patients stay temporarily in the SDA ward, typically for a few hours, and then are admitted to a




(a) Admission sources and daily admission rates (b) Patient distribution based on medical diagnosis

**Figure 3** **Four admission sources to general wards and nine medical specialties. Daily admission rates and patient distributions are estimated from Periods 1 and 2 data.**

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 3a shows the four admission sources and their average daily admission rates which are estimated from combining the Periods 1 and 2 data. Each patient is only counted once when we calculate the admission rate for the corresponding admission source, even though some patients may be transferred out of and back into GW's after the initial admission. In this paper, patients admitted to GW's from any of the four sources are called *general patients*.

## 2.2. Medical specialties

General patients are classified by one of nine medical specialities based on diagnosis at time of admission as an inpatient: Surgery, Cardiology, Orthopedic, Oncology, General Medicine, Neurology, Renal Disease, Respiratory, and Gastroenterology-Endocrine. Although Gastroenterology and Endocrine are two different medical specialties, in this paper we group them together and denote as Gastroenterology-Endocrine (Gastro-Endo or Gastro for short). 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 [51] for the same classification. We group Dental, Eye, and ENT patients into Surgery for similar reasons. As we mentioned at the beginning of Section 2, two other specialties, OG and Pediatrics are excluded from our study.

Figure 3b 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 ED, while a significant proportion of Surgery patients are EL and SDA patients.

## 2.3. Performance measure

### 2.3.1. Waiting time

Recall that we define the *waiting time* of an ED-GW patient as the duration between her bed-request time and actual admission time. In Section 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 [49, 53], except that we use the admission time to wards as the end point of the waiting period while literature usually use the time when the patient exits ED. Thus, our reported waiting time is a slight overestimation of the value computed in the conventional way. (The gap between patient exiting ED and admission to ward is about 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 ward where she originally stays, 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 about 3.5 hours for SDA patients, both longer than that of ED-GW patients (with an average less than 3 hours). The major reason could be that the ICU-GW and SDA patients have been receiving care at the current wards, so that this waiting time is not an issue unless there is a bed shortage in 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 3 in Section 6 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 non-primary 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 non-primary 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 [51]. 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 Section 5.2 of [48] for discussion on specialty-level and ward-level overflow proportions.

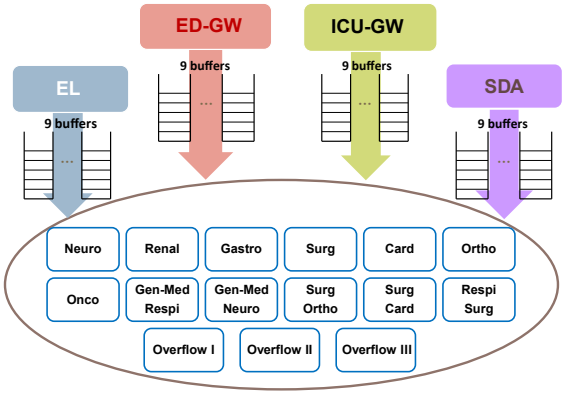



**Figure 4** **Arrival and server pool configuration in the stochastic model of NUH inpatient department.**

Besides the waiting time and overflow proportion, other performance measures of interests 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 the 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 [48] but could be adapted to other hospitals. We first give an overview of the basic ingredients of the stochastic processing network and the basic patient flow in Section 3.1. Then in Sections 3.2 to 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 patient's admission. Finally, we discuss service policies and an adjustment to incorporate patient transfer in Sections 3.5 and 3.6, respectively.

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 multi-server pools

Our proposed stochastic model is a variant of a stochastic processing network that was proposed in Harrison [24] and precisely specified in Dai and Lin [13]. 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*. Figure 4 depicts a stochastic processing network representation of the NUH inpatient department.

**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 general ward or a group of similar wards.

We use $n_j$ to denote the number servers in pool $j$, $j = 1, \ldots, J$. These $n_j$ servers are assumed to be identical. The $J$ server pools serve customers from $K$ different classes. Here, the customers are 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, following the same arrival process, service time specification, and service priority.

**Buffers.** In our model, each admission source is associated with an arrival process, which is used to model the patient bed-request process. In the rest of this paper, we use patient and customer, bed-request and arrival, and bed and server interchangeably. Each arriving patient (from any of the admission sources) is assigned to a specialty with a certain probability that depends both on the source and the arrival hour. Each arriving patient is held in a buffer, waiting to be assigned a bed and later to be admitted into the bed. The patients waiting in these buffers are processed following certain priorities which are specified by a *service policy*.

**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 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 she is *right-sited*, otherwise, *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 non-primary 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 her waiting time exceeds her pre-assigned overflow trigger time. The service policy also dictates the choice of the overflow trigger time for each patient.

**Basic patient flow.** After a bed is assigned to a patient, she has to experience extra delays (pre- and post-allocation delays) before she can be admitted to the bed. Thus, a patient's admission time is different from her bed assignment time in our model. Once a patient is admitted, she 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 corresponds to their discharge times from the hospital, and we use departure and discharge interchangeably in the rest of the paper.

### 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:

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

We will discuss the rationale for using service time model (1) in Section 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 $h_{\text{dis}}$ and $h_{\text{adm}}$ 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 12pm (noon) being .5 day. For a patient who is discharged on the same day of admission, our definition of her LOS is equal to 0, whereas when hospitals report occupancy level or some other statistics [10, 21], 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 [48], we make the following assumptions for the service time model in (1):

(a) The discharge hour $h_{\text{dis}}$ is independent of LOS and of $h_{\text{adm}}$; Section 8.5 of [48] 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 (Figure 8) and similar findings in [2].

(d) The discharge hours $h_{\text{dis}}$ for each class of patients forms another sequence of iid random variables following a certain discharge distribution. See Figure 2a in Section 1 for an example of NUH's discharge distribution.

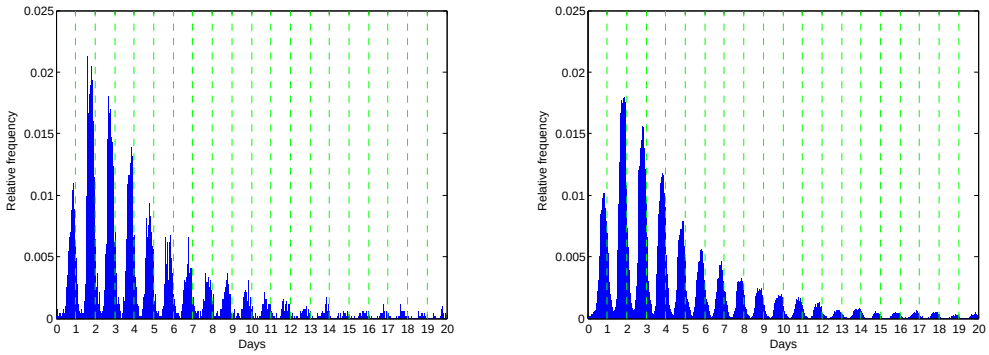
(e) We assume all iid sequences of LOS and $h_{\text{dis}}$ are independent of each other, i.e., there is no dependency among classes.

Note that for a class of patients, their admission hours $h_{\text{adm}}$ are ordered and thus cannot be iid. Though the LOS and $h_{\text{dis}}$ 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 her 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 $h_{\text{dis}}$ 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 $h_{\text{dis}}$ 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 [7]. In Section 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, while $(h_{\text{dis}} - h_{\text{adm}})$ is in the order of hours. Indeed, we can observe these two time scales from Figures 5a, which plots the empirical service time distribution at hourly resolution. On the one hand, the distribution peaks at integer values representing 1, 2, 3, ... 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 5b 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 [2, 39, 45] and appointment scheduling [3].




(a) Empirical service time distribution for Period 1; each green dashed line corresponds to a 24-hour increment

(b) Service time distribution from simulation output; LOS and discharge distributions in the simulation are empirically estimated from Period 1 data

**Figure 5 Service time distributions, at hourly resolution, for General Medicine patients that are admitted in afternoons.**

### 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.

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 which primary pool to select from.

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 her waiting time reaches $T$. After that, the patient can be assigned to either a primary bed or an overflow bed, whichever becomes available first.

#### 3.3.1. Queueing implication and QED regime

Patients can be overflowed to a non-primary server pool only if her waiting time exceeds the trigger time $T$. When $T$ is not 0, a bed can be idle even if a patient from a non-primary specialty has been waiting. Therefore, in our model the overflow policies are in general *idling*, which is different from the non-idling policies employed in many existing queueing models [33].

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 shows 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 few hours) is a small fraction of the average service time (in the order of days) and the bed utilization is high, say, $> 90\%$. 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 be achieved by a small number of servers (e.g., 30 beds in a ward).

### 3.4. Critical feature 3: allocation delays

We explicitly model operational delays that are caused by resource constraints (e.g., ED and ward nurses) other than bed unavailability during the ED to wards transfer process. Each patient in the model, even if a primary bed is available for her upon arrival, has to experience a *pre-allocation delay* first, and then a *post-allocation delay* before being admitted to the bed. We first describe the process flow from a patient's bed-request to her admission to a bed in our model, and then explain the rationale of modeling the two allocation delays. Figures 6 illustrates the process with two allocation delays under various scenarios.

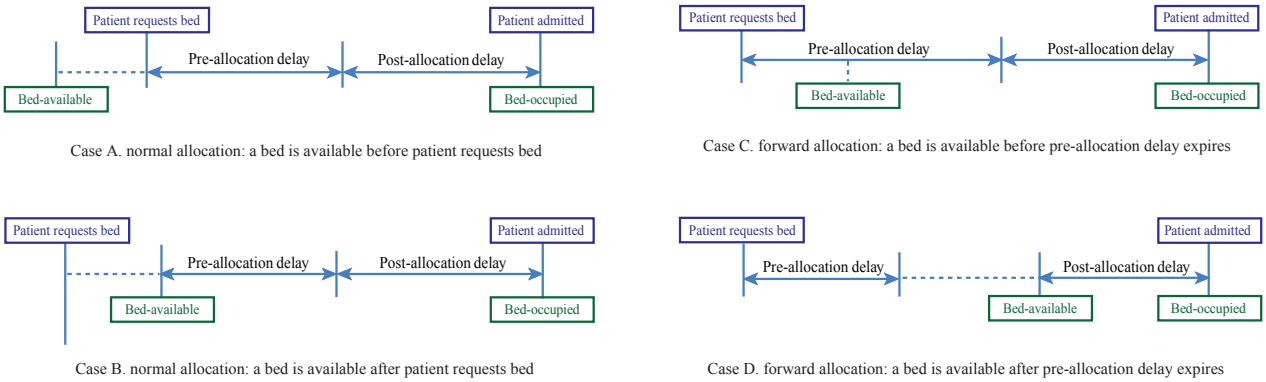



**Figure 6 Pre- and post-allocation delays under different scenarios.**

### 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 (Case A in Figure 6). If no primary bed is available, the patient waits in a buffer for a bed. When a bed becomes available and is assigned to her, following the bed assignment policy described in Section 3.3, she starts to experience a pre-allocation delay (Case B in Figure 6). 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 (Cases C and D in Figure 6). The patient starts to experience a pre-allocation delay immediately at her 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 her 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 (Case D in Figure 6). In this case, the patient waits until a bed becomes available for her, and a post-allocation delay starts at the bed-available time. When the post-allocation 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 Section 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 ED or a non-general ward and transport her 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, while 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 [48], 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 non-primary 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 (pools serving one specialty) have higher priorities than shared pools (pools serving 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 non-primary pools in 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 her. Different hospitals may adopt different overflow policies, and we will specify the time-dependent dynamic overflow policy adopted at NUH in Section 4.5.

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.

### 3.6. Modeling patient transfers between ICU and GW

In a hospital, a *real* patient can be transferred between a GW and an ICU-type ward multiple times after her initial admission to the GW. Since our proposed network has a single-pass structure, we do the following adjustments to incorporate such patient flows between GWs and ICU-type wards.

We determine an arriving patient to be a *non-transfer* or a *transfer* patient upon her arrival according to certain Bernoulli distributions. A non-transfer patient corresponds to a real patient in the hospital who does not transfer between a GW and an ICU-type ward. The transfer patient construct is used to model the first stay in a GW of a real patient who transfers to an ICU-type ward after the initial admission. Thus, the discharge (departure) time of a transfer patient in the model corresponds to the real patient's transfer-out time, and her LOS and service time are adjusted accordingly.

A real patient who transfers back to a GW after her first transfer will have a second stay in the GW. To model that second stay, we create a pseudo-patient in the model. The admission time of this pseudo-patient corresponds to the transfer-in time (from an ICU-type ward to a GW) of the modeled real patient, and the discharge time of this pseudo-patient corresponds to the final discharge time of the real patient or the next transfer-out time if the real patient transfers out of the GW again. Thus, the service time of the pseudo-patient corresponds to the duration of the second stay of the real patient. Additional pseudo-patients can be created to accommodate triple or more transfers in a similar way.

In the model, we treat the pseudo-patients as ICU-GW patients regardless of the initial admission source of the corresponding real patients. That is because the admission process and admission time

distribution of these pseudo-patients are close to those of the other ICU-GW patients according to our empirical analysis. To differentiate the two streams of ICU-GW patients, we call the pseudo-patients the *re-admitted* ICU-GW patients, and the others the *newly-admitted*. When an arrival from the ICU-GW source occurs in the model, we determine the arriving patient being newly-admitted or re-admitted according to certain Bernoulli distributions.

## 4. Populated stochastic model using NUH data

Based on the empirical study at NUH [48], we populate the proposed stochastic network model, which we refer to as the *NUH model* in the rest of the paper. 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 presents the empirical LOS and discharge distributions, while Section 4.4 introduces classification of patients based on the observations of LOS distributions. Sections 4.5 and 4.6 illustrates a dynamic overflow policy and time-varying allocation delays for the NUH model, respectively. Some remarks on the empirical study and the Companion Paper [48] are in Section 4.7.

### 4.1. Arrivals

#### 4.1.1. Time-varying arrival rates

As shown in Figure 4, patient arrivals to 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. For EL patients, their arrivals are pre-scheduled. 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.

#### 4.1.2. Arrival processes

**A time-nonhomogeneous Poisson process for ED-GW patients.** For the empirical bed-request process of ED-GW patents, we conducted a detailed study to test the assumption that it is a time-nonhomogeneous Poisson process. Following a statistical procedure proposed in [8], 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 time-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

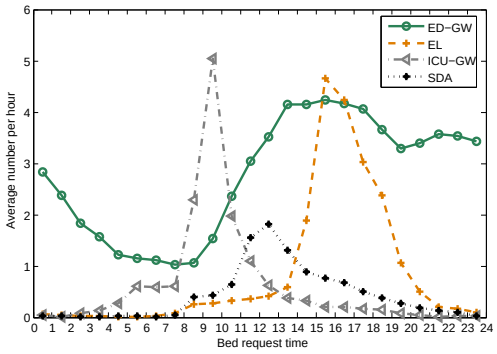



**Figure 7** **Hourly arrival rate for each admission source (estimated from Period 1 data). The daily arrival rate of each source is close to its daily admission rate shown in Figure 3a, except for the ICU-GW source since re-admitted patients are included here.**

daily arrival rate for each day of week is much higher than 1, 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 [48]. In the NUH model, we assume that the ED-GW patient's arrival process is time-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 Section 7.1.

**A non-Poisson arrival process model for other sources.** The number of EL admissions each day is pre-scheduled 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 pre-scheduled on a daily basis: ICU physicians determine the number of patients to be transferred to general wards after the morning rounds each day, and then ICU nurses submit the bed-requests for these ICU-GW patients; similar to EL patients, the SDA surgeries each day are scheduled in advance, and the SDA nurses submit bed-requests after 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 in day $k$) from admission source $j$ ($j=1,2,3$, denoting EL, ICU-GW, and SDA, respectively) at the beginning of day $k$ ($k=0,1,\dots$), 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 time-nonhomogeneous Poisson process with one day as the period [35].

### 4.2. Server pools and service policy

In the NUH model, there are 15 server pools. Table 4 in the appendix lists the number of servers and the primary specialties for each server pool.

The service policy is built based on NUH's internal guideline [40] and our empirical observations. Specifically, Table 5 in the appendix gives the priority table for components (i) and (ii) of the service policy discussed in Section 3.5. Component (iii), the overflow policy, will be elaborated in Section 4.5.

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, while 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 Section 2.1). Also see [44] 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 Section 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. Length of stay and discharge distributions

#### 4.3.1. Non-transfer patients

Table 1 lists the empirically estimated mean and standard deviation of LOS for non-transfer patients from different admission sources and specialties in the NUH model. Here, no real patients included in the empirical estimation have been transferred after admission to a GW. Transfer patients in the model have a different set of LOS distributions, and we discuss estimating their LOS distributions in Section 4.3.2.

Not surprisingly, admission source and specialty affect patient's LOS. We want to emphasize that LOS distributions are also admission-period dependent for ED-GW patients. Table 1 shows that for each specialty of ED-GW patients, a before-noon admission (AM) patient on average spends one day less than an after-noon admission (PM) patient. We speculate the reason might be that the rest of the admission day can be used for further medical diagnosis for AM patients, but not for PM 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 AM patients from these sources.

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 2a. 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.

#### 4.3.2. Transfer patients

Section 3.6 explained how to incorporate the patient flows between GW's and ICU-type wards into the model. The transfer patients we include in the NUH model are real ED-GW or EL patients

| 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) |
| Card | 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 | 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) |

**Table 1** Average LOS (in days) for non-transfer patients in each specialty from different admission sources. Period 1 data is used, and the number in the brackets is the corresponding standard deviation. Here, we list the standard deviation instead of the confidence interval to help readers fit distributions for LOS.

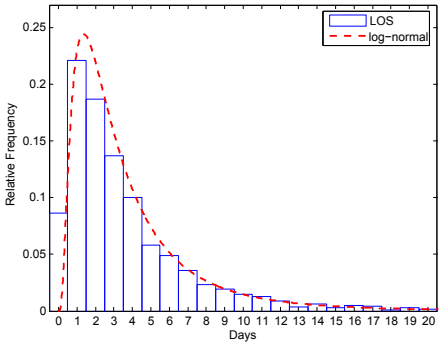



**Figure 8** **LOS of ED-GW patients from General Medicine. Only non-transfer AM patients in Period 1 are included. The LOS distribution can be fitted with a log-normal distribution (mean 3.94, std 4).**

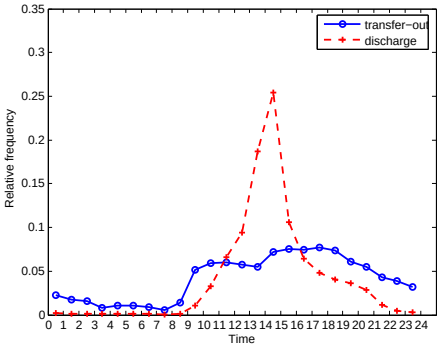



**Figure 9** **Discharge distributions for transfer patients from ED-GW and EL sources and for non-transfer patients. The solid curve is estimated from the combining Periods 1 and 2 data, and the dashed curve is estimated from Period 1 data.**

at NUH who transfer once or twice between GW's and ICU-type wards after the initial admission. We do not model (i) the real patients who are initially admitted from ICU-GW or SDA source and have been transferred; and (ii) the real ED-GW or EL patients who have transferred more than two times. We exclude them because the volume of these patients is small. Therefore, only an ED-GW or EL patient in the NUH model will be assigned to be a transfer or non-transfer type upon her arrival. An ICU-GW patient, however, will be assigned to be newly-admitted or re-admitted upon her arrival.

We use the first-visit LOS and transfer-out times of the modeled real patients to estimate the LOS distributions and discharge distributions for the transferred 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 re-admitted ICU-GW patients. The discharge distribution of the re-admitted ICU-GW patients is the same as the one for the non-transfer patients (as in Figure 2a). Figure 9 plots the discharge (transfer-out) distribution for all the transfer ED-GW and EL patients. We do not observe a significant difference between the two periods.

| $p$ | Surg | Card | 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% |

**Table 2** Estimated value for the parameter $p$ of the Bernoulli distribution to determine patient classes. 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 re-admitted 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.

### 4.4. Patient class

Patients belonging to the same class are homogeneous, having the same LOS and discharge distributions. The empirical evidence in Section 4.3 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. 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 3b 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 the LOS and discharge distributions are the same within a class.
2. Next, we determine whether (i) an ED-GW or EL patient is a non-transfer or a transfer patient, (ii) an ICU-GW patient is newly-admitted or re-admitted, following a Bernoulli distribution which 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 Sections 3.6 and 4.3.2), and are listed in Table 2.
3. Finally, at an ED-GW patient's admission time, we determine her admission period (AM or PM). By now, the patient's class is fully determined.

### 4.5. A dynamic overflow policy

At NUH, there is a general guideline [40] on when and how to overflow a patient. Consistent with this guideline, empirical evidence [51] suggests that the hospital overflows patients more aggressively during late night and early morning (before 7am). 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 prior to 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 Section 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.

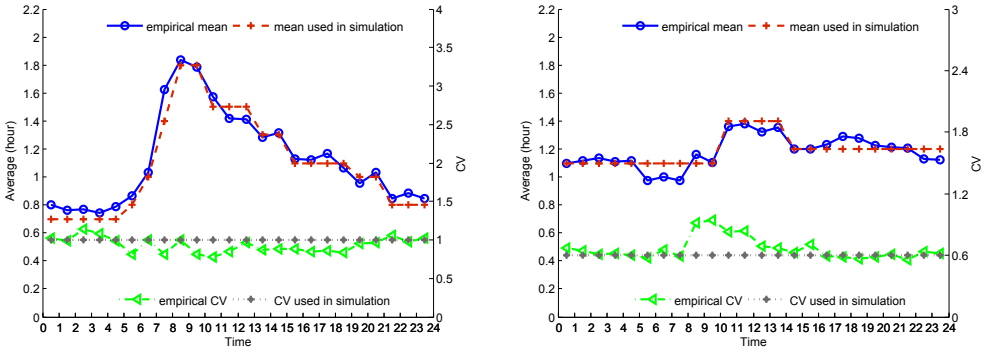
Based on these observations, we use a simple dynamic overflow policy in the NUH model: when a patient requests a bed from 7am to 7pm, the overflow trigger time $T$ is set to be $t_2 = 5.0$ hours, and for bed-requests in all other time periods, $T$ is set to be $t_1 = 0.2$ hour. We choose 7am and 7pm as the starting and ending point to adopt the long overflow trigger time, respectively. This choice is based on observations from [51] and the practice at NUH. 7pm to 7am the next day is the night-shift period at NUH. A nurse manager is in charge of dealing with all bed-requests in this period. She 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 in Figure 11 are as close to the empirical curves in the figure as possible. It is important to note that overflow decisions are very complicated [51], 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 an approximation of the real situation. Other variants of the overflow policy are possible, e.g., triggering an overflow event when the number of waiting patients exceeds a specified threshold, selecting the value of $T$ based on the remaining service times of patients who are in service. We leave this extension for future study.

### 4.6. 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 Section 4.1. For ICU-GW and SDA patients, we do not have good time stamps to estimate of their pre- and post-allocation delays reliably. We simply assume 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.

#### 4.6.1. 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.




(a) Pre-allocation delay (b) Post-allocation delay

**Figure 10** **Mean and CV of estimated pre- and post-allocation delays with respect to the delay initiation hour. Left vertical axis is for the average; 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.**

From Section 3.4.2, we know how to estimate the allocation delays empirically from the time stamps recorded in NUH data. Sub-groups 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 a 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 10a and 10b plot the empirical estimates of the mean and CV for the pre-allocation 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 8am and 10am.

The empirical curves in Figures 10a and 10b 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 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 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.

### 4.6.2. Estimating the normal allocation probabilities $p(t)$

Recall from Section 3.4 that in the model, when a patient makes a bed-request at time $t$ and there is no primary bed available at the time, we assume with probability $p(t)$ 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 do not have accurate time stamps to allow us estimate $p(t)$ reliably. In our baseline scenario, we choose

$$p(t) = \begin{cases} 0 & h(t) \in [0,6), \\ .25 & h(t) \in [6,8), \\ 1 & h(t) \in [8,12), \\ .75 & h(t) \in [12,14), \\ .5 & h(t) \in [14,20), \\ 0 & h(t) \in [20,24), \end{cases} \tag{2}$$

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 4.1 of the Online Supplement [47] 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 Section 6.4.

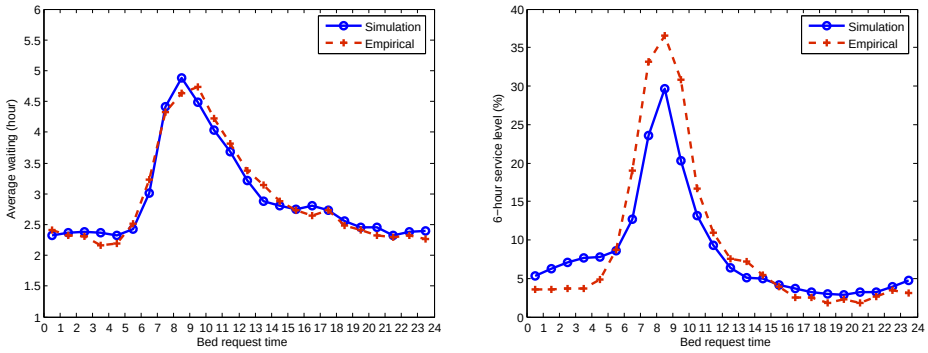
### 4.7. More empirical evidence

Due to the publication page limit, 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 [48] for the details. Specifically, Section 6 of that paper presents statistics and tests results for the arrival processes; Sections 7 and 8 document the empirical distributions for LOS and service times. Section 9 illustrates the bed-allocation and admission processes at NUH and more evidence to support our modeling of the allocation delays. Section 10 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 Sections 4.1 to 4.6, is referred to as the *baseline scenario*. In Section 5.1, we first show the simulation output from the baseline scenario matches several key empirical performance measures. Then in Section 5.2, we show that the simulation output from each model which misses one of the three critical features introduced in Section 3 cannot replicate the empirical performance measures.

To implement these models, we wrote simulation code in `C++` language. For each simulation run, we start from an empty system and simulate for a total 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 [34]. Note that in this and the next section, we rely on simulation to obtain the desired performance measures, because




(a) Hourly average waiting time (b) Hourly 6-hour service level

**Figure 11** **Baseline simulation output compares with empirical estimates: hourly average waiting time and 6-hour service level (Period 1).**

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 on-going and future analytical research in Sections 7.

### 5.1. The baseline scenario

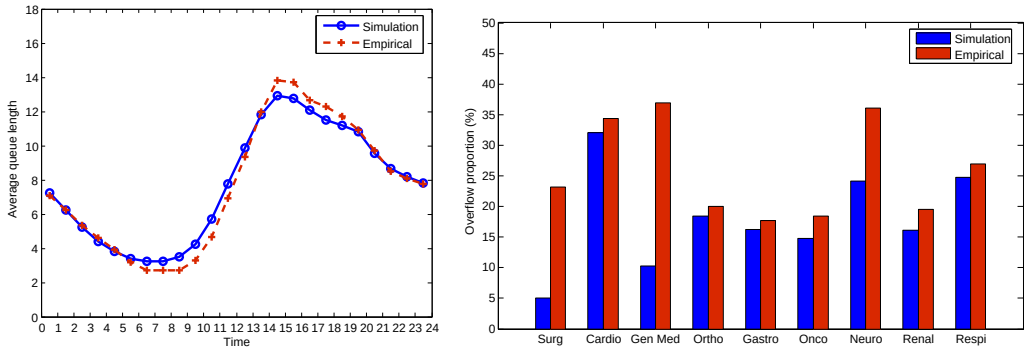
Recall that the inputs for the baseline scenario are estimated from NUH Period 1 data. Thus, we compare the outputs from this scenario against the empirical performance in Period 1 to verify the NUH model. From simulating the baseline scenario, the daily average waiting time for all ED-GW patients is 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 3 compares the simulation estimates with the empirical estimates of the average waiting time and the 6-hour service level for each specialty. 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 12a plots the hourly average queue length for all ED-GW patients for both simulation and empirical estimates.

We point out that our model cannot perfectly replicate the overflow proportion. Although the simulated overflow proportions for most specialties are close to their empirical counterparts (see Figure 12b), the baseline simulation underestimates the overflow proportions for Surgery, General Medicine, and Neurology specialties. The underestimation in these three specialties leads to an overall underestimation of overflow proportion across all specialties (16.35% in the baseline versus 26.95% from Period 1 data). Moreover, there are certain performance measures that we choose not to calibrate in the model, including the waiting time statistics for ICU-GW and SDA patients. As mentioned, the waiting time statistics for these patients are not the focus of this paper. Moreover,

| | average waiting time (hour) | | 6-h service level (%) | |
|---|---|---|---|---|
| Specialty | simulation | empirical | simulation | empirical |
| Surg | 2.64 (2.63, 2.64) | 2.61 (2.56, 2.65) | 4.85 (4.80, 4.90) | 5.45 (4.87, 6.02) |
| Card | 2.97 (2.97, 2.98) | 3.08 (3.03, 3.13) | 6.81 (6.75, 6.87) | 8.36 (7.63, 9.10) |
| Gen Med | 2.73 (2.72, 2.74) | 2.64 (2.60, 2.68) | 5.39 (5.34, 5.44) | 4.79 (4.32, 5.26) |
| Ortho | 2.73 (2.72, 2.73) | 2.79 (2.74, 2.85) | 5.22 (5.17, 5.28) | 5.84 (5.16, 6.53) |
| Gastro | 2.88 (2.88, 2.89) | 2.97 (2.90, 3.04) | 8.07 (8.00, 8.14) | 7.64 (6.73, 8.56) |
| Onco | 2.88 (2.87, 2.88) | 2.96 (2.86, 3.07) | 7.58 (7.53, 7.64) | 8.15 (6.81, 9.50) |
| Neuro | 2.84 (2.83, 2.85) | 2.81 (2.75, 2.88) | 6.49 (6.43, 6.55) | 5.93 (5.04, 6.83) |
| Renal | 3.23 (3.22, 3.24) | 3.41 (3.32, 3.51) | 10.5 (10.4, 10.5) | 11.6 (10.3, 12.9) |
| Respi | 2.82 (2.81, 2.82) | 2.77 (2.68, 2.85) | 6.25 (6.18, 6.31) | 5.50 (4.36, 6.63) |
| All | **2.82** (2.81, 2.82) | **2.82** (2.80, 2.84) | **6.29** (6.24, 6.34) | **6.52** (6.26, 6.78) |

**Table 3** Simulation and empirical estimates of waiting time statistics for ED-GW patients from each specialty. The simulation estimates are from simulating the baseline scenario, and the empirical estimates are from Period 1 data. The numbers in the parentheses are for the 95% confidence interval of the corresponding value. The confidence intervals for the simulation output are calculated following the batch mean method [34]; the confidence intervals for the empirical statistics are calculated with the standard deviations and sample sizes from the actual data.




(a) Average queue length (b) Overflow proportion

**Figure 12** **Baseline simulation output compares with empirical estimates: hourly average queue length and overflow proportion (Period 1).**

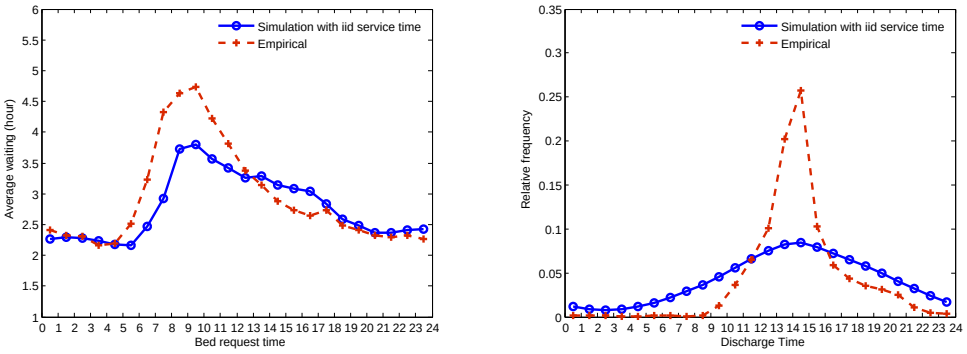
sensitivity analysis shows that whether or not we can accurately replicate their waiting times has little impact on the waiting time statistics of ED-GW patients. Readers are referred to the Online Supplement [47] for more discussion on the challenges in calibrating overflow proportions and results of sensitivity analysis.

### 5.2. Models missing any of the critical features

To show the necessity of modeling the three critical features discussed in Section 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.

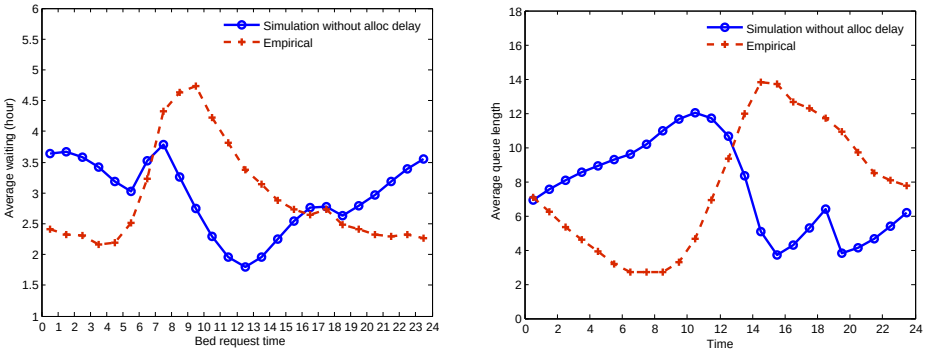
#### 5.2.1. Model with conventional iid service times

The two-time-scale service time model proposed in Section 3.2 is contrary to the exogenous, iid service time model often used in the queueing literature. We compare an iid service time model




(a) Average waiting time (b) Discharge distribution

**Figure 13 Simulation output from using an iid service time model.**



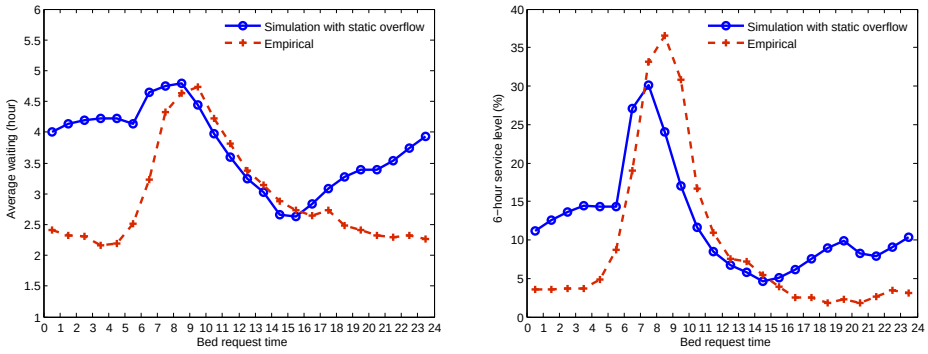

(a) Average waiting time (b) Average queue length

**Figure 14 Simulation output from a model without allocation delays.**

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. Even though this iid exogenous service time model can reproduce service time distributions such as the one in Figure 5a, it is not able to reproduce the discharge distribution and hourly waiting time statistics; 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. Section 8 of [48] contains detailed empirical observations and discussion of the iid service time model.

### 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 14b, which shows a rapid drop in the simulated average queue length between 11am and 3pm, contrasts sharply




(a) Average waiting time (b) 6-hour service level

**Figure 15 Simulation output from using a static overflow policy with a fixed overflow trigger time $T = 4.0$ hours.**

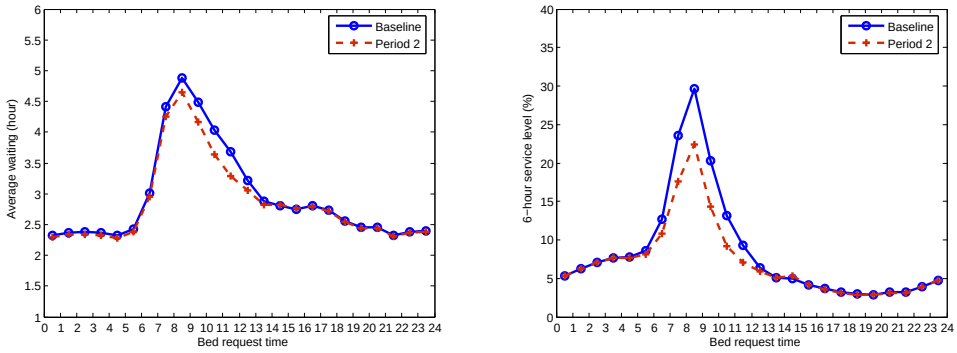
with the empirical (dashed) curve, which drops slowly after 2pm. The main reason for the rapid drop in the solid curve is that in Period 1, between 11am and 3pm, the discharge rate increases in each hour until reaching the peak at 2-3pm (see Figure 2a), and a waiting patient in the simulation is admitted 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 4 while keeping all other settings the same as 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 a 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 (10pm to 5am the next day) are about 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 even though a new primary bed is unlikely to become available within 4 hours due to the discharge pattern.

## 6. Factors that impact ED-GW patients' waiting time

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 the NUH 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 Section 6.1, we show that early discharge in Period 2 has little impact on the daily and hourly waiting time statistics. In Section 6.2, we show that a hypothetical Period 3 policy can flatten




(a) Average waiting time (b) 6-hour service level

**Figure 16** **Comparing hourly waiting time statistics under the baseline scenario and scenario with Period 2 discharge distribution.**

the hourly waiting time performance, but has limited impact on reducing the daily waiting time statistics. In Section 6.3, we study policies that mainly impact the daily waiting time performance, such as increasing bed capacity and reducing LOS. In Section 6.4, we show that most of our gained insights are robust under sensitivity analysis. Finally, we explain in Section 6.5 why these policies have different impact on the daily and hourly waiting time performance.

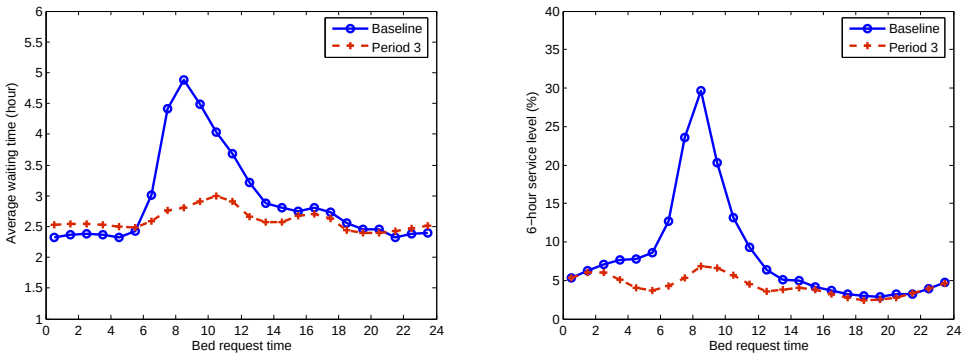
### 6.1. Period 2 discharge has a limited impact on reducing waiting time statistics

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 2a instead of the solid curve). Figure 16 compares the simulation estimates of hourly waiting time statistics with those from the baseline scenario. From Figure 16a, the hourly average waiting times show little difference between the two scenarios. From Figure 16b, the hourly 6-hour service level exhibits some reduction for bed-requests between 7am and 11am, 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 16.26%, not significantly different from the baseline value of 16.35%.

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 even though 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, which still discharges 26% of patients before noon as in Period 2, but shifts the first discharge peak time to 8-9am, i.e., three hours earlier than




(a) Average waiting time (b) 6-hour service level

**Figure 17** **Comparing hourly waiting time statistics under the baseline scenario and scenario with Period 3 policy: hypothetical discharge distribution with first peak at 8-9am and constant mean allocation delays.**

the first discharge peak time in Period 2. Figure 2b 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 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 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 a *Period 3 policy*. The Period 3 policy has not been implemented yet at NUH and may not be fully practical. We call it 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 Section 4.6) 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 (7am to noon) experience similar average waiting times (2.76 to 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 16.35% under the baseline scenario to 15.69% 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 (only implementing the aggressive early discharge policy or only stabilizing the allocation delays), the average waiting times for morning bed-requests are still about 1-2 hours longer 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. However, the Period 3 policy has limited impact on reducing the daily waiting time statistics and overflow proportion in the NUH setting.

#### 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 [47] details the tested policies and simulation results. We highlight two main observations here 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. 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 11am and noon as in Period 2, even pushing 75% 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 at 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-minutes reduction from the baseline scenario) and the overflow proportion is 14.00%.

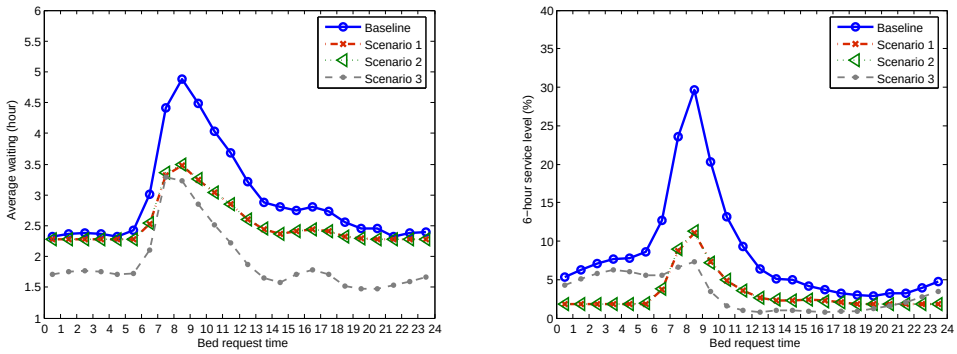
### 6.3. Policies impact on the daily waiting time statistics and overflow proportion

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; and the overflow proportions are 8.19%, 8.17%, and 15.94%, respectively. Figure 18 plots the hourly waiting time statistics under the three scenarios. Comparing to the baseline scenario, a

10% capacity increase results in a significant reduction in the overflow proportion (a 8% *absolute* 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 one hour reduction in the mean pre- and post-allocation delays leads to a one hour reduction in the daily average waiting time, while 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 the figure that in all three scenarios, the hourly average waiting time is not stabilized, i.e., the average waiting time for patients requesting beds between 7am and 11am is still about 1-2 hours longer than the daily average. The hourly 6-hour service level, though, appears to be more time-stable than the average waiting time for each scenario, especially considering that the peak value is 30% in the baseline. Note that until we increase the bed capacity to 707 beds (utilization rate reduces to 71.0%), the waiting time curves can be approximately stabilized; whereas reducing the mean allocation delays down to 0 still cannot achieve a time-stable performance.




(a) Average waiting time (b) 6-hour service level

**Figure 18** **Hourly waiting time statistics under three scenarios. 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 for each. In each scenario, we use the Period 1 discharge distribution and assume the constant-mean allocation delay model.**

In summary, our model predicts that reducing the mean allocation delays can significantly reduce the daily average waiting times, while increasing the bed capacity or reducing the LOS mainly impact the overflow proportion in the NUH setting. Moreover, these policies have less impact on the time-of-day pattern of waiting time performance and they do not necessarily flatten the hourly waiting time performance. Our results 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 that the long waiting time of ED-GW patients may not be caused by a lack of inpatient beds but rather by other inefficiencies which slow the transitions of care between different hospital units [41]. 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, and choosing different values for the normal allocation probability $p(t)$. 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 (reduce from 4.37 to about 2.5 hours); (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 AM- and PM-admitted ED-GW patients (see Table 1). Sections 2 and 3 of the Online Supplement detail all the experiment setting and simulation results of the sensitivity analysis discussed in this section.

### 6.5. Intuition about the gained insights

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 [14].

There are two types of waiting 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 till next day to get a bed. (ii) Within a day, the discharge timing is too late so that morning arrivals have to wait several hours (till the afternoon) to get a bed. The first type of waiting, reflected in the daily waiting time performance, can be affected by the daily arrival rate, LOS distributions and bed capacity. The second type of waiting, 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 waiting, not the first. Thus, early discharge policies can flatten hourly waiting time curves, but has limited impact on further reducing 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 a modest number of patients be discharged early enough (before 9-10am in the NUH setting as

suggested by our simulation results) so 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 11am, which is not early enough.

In comparison, increasing capacity or reducing LOS helps reduce the first type of waiting 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 waiting due to not enough patients being discharged early enough. This is why increasing capacity does not necessarily flatten the hourly waiting time curves.

## 7. Concluding remarks and future research

We have proposed a high-fidelity stochastic network model for inpatient flow management, which 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 delay significantly reduces the daily average waiting time, and increasing bed capacity or reducing LOS can greatly reduce the 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 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.

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 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 8-9am is difficult since physicians and nurses are busy with the morning rounds at about this time. On the other hand, stabilizing allocation delays also requires coordination throughout the entire hospital and proper staffing at

different units at various hours of the day. Though 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. Besides the discharge policy, our model allows hospital managers to evaluate the trade-off between the benefit of reducing ED overcrowding and the cost of implementing a number of operational and strategic polices.

### 7.1. Future work

Our 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, we use pre- and post-allocation delays as two black boxes to model all possible secondary bottlenecks including ward nurse and BMU staff shortage at certain times of the day, partly because we want to maintain the tractability of the proposed model. Detailed studies are needed to further understand these secondary bottlenecks so that we can explicitly incorporate them into the model and identify strategies to reduce allocation delays. The two-queue model proposed in [57] 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 [11, 31] 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 can 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 a week. We also assume that elective admissions are stationary by day, while recent work pointed out that elective schedule is actually the main source of daily occupancy variation in many hospitals [25]. Our model can be extended to predict day-of-week performance and help design a better elective schedule.

Finally, to obtain structural insights into the impact of many policies such as discharge timing and overflow trigger time, simulation 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 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 [14].

## References

[1] G. Allon, S. Deo, and W. Lin, "The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence," 2012, working paper. [Online]. Available: http://ssrn.com/abstract=1516843

[2] M. Armony, S. Israelit, A. Mandelbaum, Y. Marmor, Y. Tseytlin, and G. Yom-Tov, "Patient flow in hospitals: A data-based queueing perspective," 2011, working paper. [Online]. Available: http://www.stern.nyu.edu/om/faculty/armony/Patient%20flow%20main.pdf

[3] M. Armony and C. Zacharias, "Panel sizing and appointment scheduling in outpatient medical care," 2013, working paper.

[4] A. Bair, W. Song, Y. Chen, and B. Morris, "The impact of inpatient boarding on ED efficiency:a discrete-event simulation study," *J Med Syst*, vol. 34, pp. 919–929, 2010.

[5] D. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[6] A. Birjandi and L. M. Bragg, *Discharge Planning Handbook for Healthcare: Top 10 Secrets to Unlocking a New Revenue Pipeline*. New York: Productivity Press, 2008.

[7] I. Borghans, R. Heijink, T. Kool, R. Lagoe, and G. Westert, "Benchmarking and reducing length of stay in Dutch hospitals," *BMC Health Services Research*, vol. 8, no. 1, p. 220, 2008.

[8] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao, "Statistical analysis of a telephone call center," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 36–50, 2005.

[9] A. M. d. Bruin, A. v. Rossum, M. C. Visser, and G. Koole, "Modeling the emergency cardiac in-patient flow: An application of queuing theory," *Health Care Management Science*, pp. 1–13, 2006.

[10] Centers for Disease Control and Prevention, USA, "Health, United States," 2010. [Online]. Available: http://www.cdc.gov/nchs/data/hus/hus10.pdf

[11] C. Chan, G. Yom-Tov, and G. J. Escobar, "When to use Speedup: An Examination of Intensive Care Units with Readmissions," 2011, working paper. [Online]. Available: http://www.columbia.edu/~cc3179/ICU_fluid.pdf

[12] J. Cochran and A. Bharti, "Stochastic bed balancing of an obstetrics hospital," *Health Care Management Science*, vol. 9, no. 1, pp. 31–45, 2006.

[13] J. G. Dai and W. Lin, "Maximum pressure policies in stochastic processing networks," *Operations Research*, vol. 53, pp. 197–218, 2005.

[14] J. G. Dai and P. Shi, "A two-time-scale approach to time-varying queues for hospital inpatient flow management," 2012, working paper.

[15] Department of Health, United Kingdom, "Achieving timely simple discharge from hospital: A toolkit for the multi-disciplinary team," 2004. [Online]. Available: http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4088366

[16] Z. Feldman, A. Mandelbaum, W. A. Massey, and W. Whitt, "Staffing of time-varying queues to achieve time-stable performance," *Management Science*, vol. 54, no. 2, pp. 324–338, 2008.

[17] N. Gans, G. Koole, and A. Mandelbaum, "Telephone call centers: Tutorial, review, and research prospects," *Manufacturing & Service Operations Management*, vol. 5, no. 2, pp. 79–141, 2003.

[18] L. Green, "Queueing analysis in healthcare," in *Patient Flow: Reducing Delay in Healthcare Delivery*, ser. International Series in Operations Research and Management Science, R. W. Hall, Ed. Springer US, 2006, vol. 91, pp. 281–307.

[19] ——, "How many hospital beds?" *Inquiry*, vol. 39, no. 4, pp. 400–412, 2002.

[20] J. Griffin, S. Xia, S. Peng, and P. Keskinocak, "Improving patient flow in an obstetric unit," *Health Care Manag Sci*, 2011.

[21] M. J. Hall, C. J. DeFrances, S. N. Williams, A. Golosinskiy, and A. Schwartzman, "National hospital discharge survey: 2007 summary," *Natl Health Stat Report*, no. 29, pp. 1–20, 24, 2010.

[22] R. Hall, "Bed assignment and bed management," in *Handbook of Healthcare System Scheduling*, ser. International Series in Operations Research & Management Science, R. Hall, Ed. Springer US, 2012, vol. 168, pp. 177–200.

[23] R. Hall, D. Belson, P. Murali, and M. Dessouky, "Modeling patient flows through the healthcare system," in *Patient Flow: Reducing Delay in Healthcare Delivery*, R. Hall, Ed. Springer, 2006.

[24] J. M. Harrison, "Brownian models of open processing networks: canonical representation of workload," *Annals of Applied Probability*, vol. 10, pp. 75–103, 2000, correction: **13**, 390–393 (2003).

[25] J. Helm and M. Van Oyen, "Design and optimization methods for elective hospital admissions," 2012, working paper.

[26] J. E. Helm, S. AhmadBeygi, and M. P. Van Oyen, "Design and analysis of hospital admission control for operational effectiveness," *Production and Operations Management*, vol. 20, no. 3, pp. 359–374, 2011.

[27] N. Hoot and D. Aronsky, "Systematic review of emergency department crowding: Causes, effects, and solutions." *Ann Emerg Med*, vol. 52, pp. 126–36, 2008.

[28] E. Howell, E. Bessman, S. Kravet, K. Kolodner, R. Marshall, and S. Wright, "Active bed management by hospitalists and emergency department throughput." *Annals of Internal Medicine*, vol. 149, no. 11, pp. 804–810, 2008.

[29] S. H. Jacobson, S. N. Hall, and J. R. Swisher, "Discrete-event simulation of health care systems," in *Patient Flow: Reducing Delay in Healthcare Delivery*, ser. International Series in Operations Research and Management Science, R. W. Hall, Ed. Springer US, 2006, vol. 91, pp. 211–252.

[30] S. Khanna, J. Boyle, N. Good, and J. Lind, "Impact of admission and discharge peak times on hospital overcrowding," *Health Informatics: The Transformative Power of Innovation*, pp. 82–88, 2011.

[31] S.-H. Kim, C. Chan, M. Olivares, and G. J. Escobar, "ICU admission control: An empirical study of capacity allocation and its implication on patient outcomes," 2012, working paper. [Online]. Available: http://www.columbia.edu/~cc3179/ICUadm_2012.pdf

[32] N. Koizumi, E. Kuno, and T. E. Smith, "Modeling patient flows using a queuing network with blocking," *Health care management science*, vol. 8, no. 1, pp. 49–60, 2005.

[33] P. R. Kumar, "Re-entrant lines," *Queueing Systems*, vol. 13, pp. 87–110, 1993.

[34] A. M. Law and D. W. Kelton, *Simulation Modelling and Analysis*. McGraw-Hill Education - Europe, 2000.

[35] P. A. Lewis and G. S. Shedler, "Simulation methods for Poisson processes in nonstationary systems," in *Proceedings of the 10th conference on Winter simulation - Volume 1*, ser. WSC '78. Piscataway, NJ, USA: IEEE Press, 1978, pp. 155–163.

[36] E. Litvak, M. C. Long, A. B. Cooper, and M. L. McManus, "Emergency department diversion: Causes and solutions," *Academic Emergency Medicine*, vol. 8, no. 11, pp. 1108–110, 2001.

[37] S. W. Liu, S. H. Thomas, J. A. Gordon, A. G. Hamedani, and J. S. Weissman, "A pilot study examining undesirable events among emergency department-boarded patients awaiting inpatient beds," *Annals of Emergency Medicine*, vol. 54, no. 3, pp. 381–385, 2009.

[38] Y. Liu and W. Whitt, "Stabilizing customer abandonment in many-server queues with time-varying arrivals," 2012, working paper. [Online]. Available: http://www.columbia.edu/~ww2040/LiuWhittStabilizing032812.pdf

[39] A. Mandelbaum, P. Momcilovic, and Y. Tseytlin, "On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers," *Management Science*, 2012.

[40] National University Hospital, "BMU training guide: Inpatient operations," December 2011.

[41] J. M. Pines, R. J. Batt, J. A. Hilton, and C. Terwiesch, "The financial consequences of lost demand and reducing boarding in hospital emergency departments," *Annals of Emergency Medicine*, vol. 58, no. 4, pp. 331–340, 2011.

[42] J. M. Pines, J. A. Hilton, E. J. Weber, A. J. Alkemade, H. Al Shabanah, P. D. Anderson, M. Bernhard, A. Bertini, A. Gries, S. Ferrandiz, V. A. Kumar, V.-P. Harjola, B. Hogan, B. Madsen, S. Mason, G. Ohlen, T. Rainer, N. Rathlev, E. Revue, D. Richardson, M. Sattarian, and M. J. Schull, "International perspectives on emergency department crowding," *Academic Emergency Medicine*, vol. 18, no. 12, pp. 1358–1370, 2011.

[43] J. M. Pines, S. Iyer, M. Disbot, J. E. Hollander, F. S. Shofer, and E. M. Datner, "The effect of emergency department crowding on patient satisfaction for admitted patients," *Academic Emergency Medicine*, vol. 15, no. 9, pp. 825–831, 2008.

[44] E. S. Powell, R. K. Khare, A. K. Venkatesh, B. D. Van Roo, J. G. Adams, and G. Reinhardt, "The relationship between inpatient discharge timing and emergency department boarding," *The Journal of Emergency Medicine*, 2011.

[45] M. Ramakrishnan, D. Sier, and P. Taylor, "A two-time-scale model for hospital patient flow," *IMA Journal of Management Mathematics*, vol. 16, no. 3, pp. 197–215, 2005.

[46] S. Schneider, F. Zwemer, A. Doniger, R. Dick, T. Czapranski, and E. Davis, "Rochester, New York: a decade of emergency department overcrowding," *Academic Emergency Medicine*, vol. 8, no. 11, pp. 1044–1050, 2001.

[47] P. Shi, M. Chou, J. G. Dai, D. Ding, and J. Sim, "Online Supplement for "Models and Insights for Hospital Inpatient Operations: Time-of-Day Congestion for ED Patients Awaiting Beds"," 2012, online supplement.

[48] P. Shi, J. G. Dai, D. Ding, J. Ang, M. Chou, J. Xin, and J. Sim, "Patient Flow from Emergency Department to Inpatient Wards: Empirical Observations from a Singaporean Hospital," 2012.

[49] Sinagpore Ministry of Health, "Waiting time for admission to ward," 2012. [Online]. Available: http://www.moh.gov.sg/content/moh_web/home/statistics/healthcare_institutionstatistics/Waiting_Time_for_Admission_to_Ward.html

[50] A. J. Singer, J. Thode, Henry C., P. Viccellio, and J. M. Pines, "The association between length of emergency department boarding and mortality," *Academic Emergency Medicine*, vol. 18, no. 12, pp. 1324–1329, 2011.

[51] K. Teow, E. El-Darzi, C. Foo, X. Jin, and J. Sim, "Intelligent analysis of acute bed overflow in a tertiary hospital in Singapore," *Journal of Medical Systems*, pp. 1–10, Janurary 2011.

[52] S. Thompson, M. Nunez, R. Garfinkel, and M. Dean, "Efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges," *Operations Research*, vol. 57, no. 2, pp. 261 – 273, 2009.

[53] United States General Accounting Office, *Hospital emergency departments: crowded conditions vary among hospitals and communities*. Washington, D.C.: United States General Accounting Office, 2003.

[54] F. d. Vericourt and O. B. Jennings, "Nurse staffing in medical units: A queueing perspective," *Operations Research*, vol. 59, no. 6, pp. 1320–1331, 2011.

[55] H. J. Wong, D. Morra, M. Caesar, M. W. Carter, and H. Abrams, "Understanding hospital and emergency department congestion: An examination of inpatient admission trends and bed resources," *Canadian Journal of Emergency Medicine*, vol. 34, no. 1, pp. 18–26, 2010.

[56] D. A. Yancer, D. Foshee, H. Cole, R. Beauchamp, W. de la Pena, T. Keefe, W. Smith, K. Zimmerman, M. Lavine, and B. Toops, "Managing capacity to reduce emergency department overcrowding and ambulance diversions," *Jt Comm J Qual Patient Saf*, vol. 32, no. 5, pp. 239–45, 2006.

[57] N. Yankovic and L. V. Green, "Identifying good nursing levels: A queuing approach," *Operations Research*, vol. 59, no. 4, pp. 942–955, 2011.

[58] D. D. Yao, *Stochastic Modeling and Analysis of Manufacturing Systems*, ser. Springer Series in Operations Research. New York: Springer, 1994.

[59] S. Zeltyn, Y. N. Marmor, A. Mandelbaum, B. Carmeli, O. Greenshpan, Y. Mesika, S. Wasserkrug, P. Vortman, A. Shtub, T. Lauterman, D. Schwartz, K. Moskovitch, S. Tzafrir, and F. Basis, "Simulation-based models of emergency departments: Operational, tactical, and strategic staffing," *ACM Trans. Model. Comput. Simul.*, vol. 21, no. 4, pp. 24:1–24:25, Sep. 2011.

## Appendix. Server pool setting and service policy in the NUH model

This appendix presents Tables 4 and 5 that are needed to complete the specification of server pools and service policy in Section 4.2.

Table 4 lists 15 server pools. Each row specifies one server pool with the basic information including index, primary specialty, and number of servers. The table is based on the empirical study at NUH. We slightly adjust the number of servers in certain server pools because our proposed stochastic model does not capture all the constraints in bed assignment. For example, Orthopedic patients with open wounds cannot stay in the same room with patients who have acquired Methicillin-resistant Staphylococcus Aureus (MRSA), while our model does not differentiate MRSA patients from non-MRSA patients. Moreover, our model does not explicitly consider patients' preference for bed classes (beds in private and shared rooms are in different classes; see [48] for details of bed classes.) To compensate for the inefficiency caused by class mismatch in the real hospital setting,

| 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 | Card | 40 |
| 7 | Ortho | 50 |
| 8 | Onco | 43 |
| 9 | Respi, Surg | 25 |
| 10 | Surg, Ortho | 38 |
| 11 | Surg, Card | 30 |
| 12 | Overflow ward I | 39 |
| 13 | Overflow ward II | 43 |
| 14 | Overflow ward III | 48 |
| Total | | 563 |

**Table 4** Server pool index, primary specialty, and number of servers.

| 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 |

**Table 5** Priority of primary and overflow pools; pool numbers are ordered in decreasing priority.

we assume pools 12, 13, and 14, which correspond to three wards that have class A or class B1 beds, to be overflow pools. These three pools only accept patients whose overflow trigger times are reached in the model. This adjustment is based on the facts that these wards usually do not admit patients who prefer class B2 or class C beds (for financial reasons) except for urgent situations.

For components (i) and (ii) in the service policy, Table 5 specifies the priority of primary pools for arrivals, and the priority of non-primary pools when overflowing a patient. This table is constructed based on empirical studies, NUH's internal bed allocation guideline [40], and discussions with BMU staff.