Online Supplement for "Hospital Inpatient Operations: Mathematical Models and Managerial Insights"

Pengyi Shi, J. G. Dai
School of Industrial and Systems Engineering
Georgia Institute of Technology, Atlanta, GA 30332
{pengyishi,dai}@gatech.edu

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

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

Xin Jin, Joe Sim
National University Hospital, Singapore
{xin_jin,joe_sim}@nuhs.edu.sg

August 22, 2012

Contents

1 Introduction and outline

This document is an online supplement to the main paper [25], which proposes a novel stochastic network model to capture the inpatient operations in a Singaporean hospital and to understand the effect of an early discharge policy on waiting time for admission to wards. This supplement presents a comprehensive empirical study on the inpatient flow management in this hospital with data gathered from 2008 to 2010, providing a basis to construct the proposed model. We report the statistics of waiting times for patients admitted from the emergence department (ED) to inpatient wards (referred to as ED-GW patients in this document and in [25]), bed occupancy rate (BOR) of each ward, and overflow proportion for each ward. We also report the probability distributions and parameters related to arrival, discharge, length of stay (LOS), and pre- and post-allocation delays. These empirical results generate the inputs for probability distributions and parameters that are needed for the simulation analysis of the stochastic model proposed in [25].

We also hope that the empirical results documented in this online supplement will provide fresh insights for researchers interested in stochastic networks, patient flow models, and inpatient department operations. For example, an operational insight we achieved is that the LOS for patients admitted before noon is statistically more than one day shorter than for patients admitted after noon. Besides serving as an online supplement to the main paper, this document can potentially stimulate future research in stochastic modeling of patient flow management.

Outline

This online supplement is organized as follows. Section 2 gives an overview of the inpatient department in this Singaporean hospital. Section 3 describes an early discharge campaign implemented in this hospital, and introduces the reason of using two periods (Period 1 and 2) in the empirical analysis. Sections 4 and 5 introduce two key performance measures, the waiting time of ED-GW patients and the overflow proportion. The statistics on these two performances are reported. Section 5 also describes the basic organizational units in the hospital, wards, and reports ward-level statistics such as BOR. Sections 6 to 9 relate to the modeling elements of the proposed stochastic network model in [25]. Section 6 discusses the bed-request process from ED-GW patients (which serves as the arrival process to the stochastic model). Sections 7 and 8 are for the service time model. Section 9 summarizes the motivation of modeling the allocation delays and relevant empirical studies. Section 10 presents a detailed study for all transfer patients, in particular, for the transfer patients who are not captured in the stochastic model in [25]. Finally, Section 11 specifies some additional details of the simulation input and shows simulation results that are not included in Section 5 of the main paper.

2 NUH Inpatient department

National University Hospital (NUH) is one of the major public hospitals in Singapore. It operates a busy emergency department (ED) and a large inpatient department that has about 1000 beds as of January 1, 2011. In this section, we describe the inpatient department's operations. We first define the *general wards* that will be the focus of this study, and the four admission sources for patients admitted to these wards.

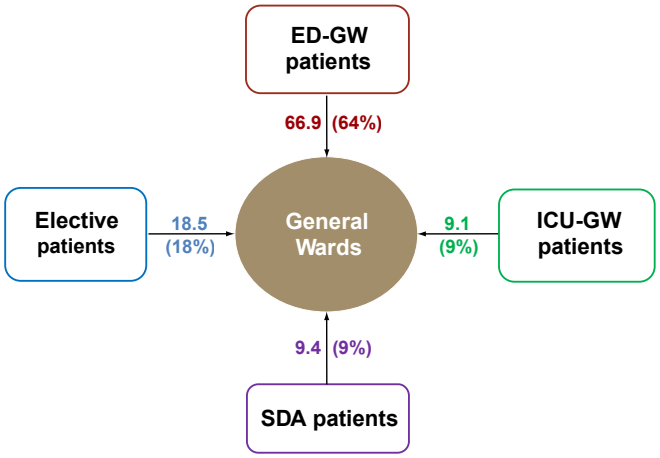


Figure 1: Admission sources to general wards and their daily admission rates

2.1 General wards

In this study, we focus on 19 general wards (GWs) having a total 555 to 638 beds between January 1, 2008 and December 31, 2010. These wards exclude intensive-care-unit (ICU) wards (55-67 beds), isolation wards (42-44 beds), and high-dependence wards (71-103 beds). The 16 beds in the extended-diagnostic-treatment-unit (ETDU) (an observational unit for patients from the emergency department (ED)), and about 50 beds in the same-day-admission (SDA) unit are excluded. The maternity wards and pediatric wards are also excluded. See Section 2.4 for a discussion on their exclusion. Table 5 in Section 5 specifies the 19 general wards.

2.2 Admission sources

We categorize inpatients admitted to GWs into four categories according to their admission sources. They are ED-GW, ICU-GW, Elective (EL), and SDA patients. We will elaborate on each of these categories below. Admissions from Period 1 (18 months) account for 55357 inpatients, with a daily admission rate of 101.20, and those from Period 2 (12 months) account for 39429 inpatients, with a daily admission rate of 108.02. Figure 1 depicts the four admission sources and gives the average daily admission rate of each source. These rates are estimated from the combined data set. When we calculate the daily admission rate (overall and for each admission source), each patient with initial admission from the corresponding source is counted once, even though some of them go through a sequence of transfers later, such as from GW to ICU, and then back to GW. In other words, we only count the initial admissions to GW and do not double count the transfers. Section 10 will elaborate the details of internal transfer after initial admission. As explained in Section 3.4 of the main paper [25], the transfers of certain patients are modeled as a steam of pseudo-arrivals (i.e., re-admitted class under ICU-GW source) in the proposed stochastic model.

ED-GW patients

The emergency department (ED) services provide treatment to patients in need of urgent medical care, and determine the timely transition to the next stage of definitive care, if necessary. Of the 310519 patients who visit ED in our combined data set (from either ambulance or walk-in arrivals),

61018 (19.7%) patients are admitted to the GWs and become ED-GW patients. 213078 (68.6%) patients are treated and directly discharged from ED because of death, absconded, admission no show, transferred to other hospital, followed up at Specialist Outpatient Clinic (SOC), Primary Health Care (PHC), General Practitioner (GP), and discharges to home. The remainder are admitted to an ICU-type (ICU, isolation, or high-dependency) ward (12163 patients, 3.9%) for further medical care, or to the EDTU (10180 patients, 3.3%) for further observation, or to other wards such as the Endoscopy ward.

The waiting time of an ED-GW patient, from the time of bed-request to time of leaving the ED, is a key performance measure. The Ministry of Health (MOH) uses the phrase, "waiting time for admission to ward" for this performance measure [27] and monitors it closely. MOH requires weekly reports on certain statistics of this waiting time from every public hospital in Singapore. In the medical literature, the waiting time of ED-GW patients is also known as the "ED boarding time" [30]. See Section 4 for a detailed discussion of waiting time statistics for ED-GW patients.

ICU-GW patients

Of the 13988 patients initially admitted to ICU-type wards (from either ED or other admission sources) in our combined data set, 8282 (59.2%) of them transfer to GWs later. These patients are labeled ICU-GW patients. The others are discharged directly from an ICU-type ward. The delay between bed-request and departure from an ICU-type ward for ICU-GW patients is considered less important than the waiting time of ED-GW patients, because ICU-GW patients are receiving satisfactory care in the ICU-type ward. This waiting time is only important when there is a shortage of ICU-type beds.

Elective patients

Elective (EL) patients usually have less urgent medical conditions than ED-GW patients. They are referred by clinical physicians and are admitted to the hospital through "elective referrals" by the hospital specialists. Most of the EL patients come to seek surgeries, and they are admitted at least one day prior to surgery.

The daily number of admissions from EL patients are pre-scheduled (see for example [12]). The beds for these scheduled patients are usually reserved so that patients need not wait for their beds when they arrive at the hospital. Moreover, the arrival times of EL patients (the time when presenting at wards) are also scheduled as the patients are typically advised to come in the afternoon. As a result, there is no meaningful time stamp for EL patient's bed-request time as for patients from other sources. In [25], the authors use the empirical admission time to model the bed-request time of EL patients in the stochastic model, and give them the highest priority to ensure the gap between the bed-request time and admission time is negligible.

SDA patients

Same-day-admission (SDA) patients first go to the operating rooms for surgical procedures, usually in the morning, occupy a temporary bed until recovery, and are finally admitted to a GW. An SDA patient is similar to an EL patient except that the EL patient is admitted into a GW *before* the day of surgery, whereas the SDA patient is admitted to a GW *after* the surgery. Therefore, it is expected that an EL patient typically stays in a general ward bed at least one day longer than an SDA patient.

Besides the four admission sources we described above, there are a few patients (around 2.5%), who are admitted to general wards from other sources. For example, some patients are transferred

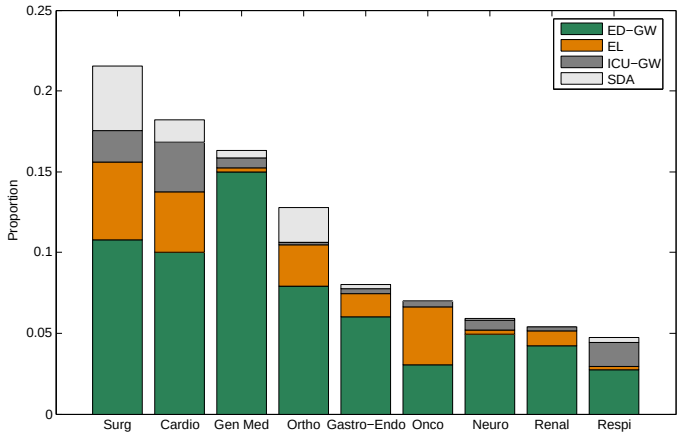


Figure 2: General ward patient distribution via medical diagnosis

from EDTU or Endoscopy ward to GW. In this document as well as in [25], we lump these patients into the SDA admission source due to their similar admission patterns and length of stay (LOS) distributions. In Figure 1, the daily admission rate for "SDA patients" already includes these patients.

2.3 Medical specialties

At admission, NUH categorizes adult inpatients into ten major specialities based on their diagnosis: Surgery, Cardiology, Orthopedic, Oncology, General Medicine, Neurology, Renal disease, Respiratory, Gastroenterology-Endocrine, and Obstetrics and Gynecology (OG). Although Gastroenterology and Endocrine are two different medical specialties, this online supplement and the main paper [25] group them under the name of Gastroenterology-Endocrine (Gastro-Endo or Gastro for short), because patients in these two specialties always share the same ward and have a similar LOS distribution. See the same aggregation in [29]. Dental, Eye, and Otolaryngology (ENT) are grouped under Surgery for a similar reason. As explained in Section 2.1, we exclude OG from our study; see Section 2.4 for discussion.

Figure 2 shows the patient distribution according to the medical specialty. Note that NUH also uses the term "cluster" to classify its inpatients. The major clusters are Medicine, Surgical, Cardiology, Orthopedic, Oncology, and OG. Within a cluster, patients are further classified by medical discipline, e.g., the Medicine cluster includes the five specialties we mentioned above: General Medicine, Gastro-Endo, Neurology, Renal disease, and Respiratory (there are about 2% of Medicine patients belonging to minor disciplines such as Rheumatology and Geriatrics; we group them under Gastro-Endo for convenience). Although a cluster is an important organization unit within NUH, it is not an essential concept in this study. For ease of exposition, we choose to use a generic name "specialty" to classify patients with various medical diagnoses.

2.4 Rationales for excluding certain wards

The entire inpatient department has 38 wards in total. As mentioned, we exclude 13 special care units from our defined general wards, i.e., 5 ICU wards, 5 high-dependency units (HD), 2 isolation units (ISO), and a delivery ward. It is because these wards are dedicated to patients with special needs and therefore have different performance expectations. We call ICU, HD, and ISO wards *ICU-type wards*, and consider the interactions between GW and ICU-type wards through ICU-GW patients in the model (see Section 3.4 of [25] for details).

We exclude four Pediatric wards because they act independently from the rest of the hospital. The hospital rarely assigns an adult inpatient to a Pediatric ward (1% incidence), and Pediatric inpatients rarely stays in adult wards (0.8% incidence). Moreover, the hospital has a dedicated children's emergency department with its own admission process and a Pediatric intensive care unit (PICU) for critically ill newborns and children. Thus, Pediatric patients have few interactions with adult patients, and their performances are not the focus of our study.

Finally, we exclude two OG wards for a similar reason. Less than 1% OG patients stay in non-OG wards, and less than 0.5% non-OG adult patients are admitted to OG wards. Moreover, OG patients have very different admission patterns from other adult patients. Most of them come to deliver babies, so they go to the delivery ward or SDA ward first, and then transfer to OG wards; a few of them are directly admitted from ED. Their length of stay (LOS) in the hospital is also significantly shorter than other patients. If they are included, then many of our reported performance measures will be distorted.

In summary, we focus on the remaining 19 "general wards" in the empirical study. See Table 5 in Section 5.1 for a list of these wards. We refer the inpatient beds in these wards as "general beds". We exclude all patients from our analysis who do not use general beds. Moreover, the very few OG and Pediatric patients who stay in general wards are also excluded. We refer to the remaining patients admitted to general wards as "general patients" and categorize them by the four admission sources and the nine specialties as introduced in Sections 2.2 and 2.3. All the empirical results reported in this document are based on these general patients.

2.5 Data sets

We obtain four data sets from NUH, i.e., admission data, discharge data, emergency attendance data and internal transfer data. Each of the data sets contains corresponding data entries from January 1, 2008 to December 31, 2010. We merge the four data sets using patient ID and case number as identifiers. Each record in the merged data contains a patient's entire inpatient care history. The admission related information includes admission date and time, allocated ward and bed, specialties, etc. The discharge related information includes discharge date and time, discharged ward, diagnostic code, etc. Based on whether a match in the ED attendance data set can be found, we classify each record as "visited ED" or "No visit to ED". For a patient visited ED, the ED related information includes "Trauma Start" time (time of inpatient bed request) and "Trauma End" time (time of leaving ED), etc. Certain patients went through one or more internal transfers during their hospital stays (between initial admission and final discharge). We label these patients with matched record(s) in the internal transfer data as "having been transferred". The transfer related information includes transfer frequency, transfer in and out time, target ward and bed, etc.

Records with admission or discharge time outside the three-year period have incomplete admission or discharge information. For example, if a patient is admitted before January 1, 2008 and is discharged on January 15, 2008, then her admission information is missing. We note that there are 650 records lacking admission information and 10 lacking discharge information. To ensure consistency, we apply the following conventions to these records:

1. All records with complete admission information are included in any analysis that is related to inpatient admission (e.g., admission time, daily admission rate), no matter whether discharge information is missing or not.

2. All records with complete discharge information are included in any analysis that is related to discharge (e.g., discharge distribution in Section 3), no matter whether admission information is missing or not.

3. Only records with both admission and discharge information are included in the analysis for LOS and service time (e.g., Sections 7 and 8). Records with either missing admission or discharge information are excluded.

4. All records identified as "visited ED" and with complete admission information are included in analyzing ED-GW patient's waiting times and bed-request rates, no matter whether discharge information is missing or not.

As a consequence of following these conventions, the total sample size varies for different analyses.

Extra data set on bed request information

In the main paper [25], the authors introduce a key modeling element, pre- and post-allocation delays to model inpatient operations. See Section 4.1 of the main paper and Section 9 of this document for details. To empirically estimate the distributions of the pre- and post-allocation delays, we obtain an extra data set with a different set of time stamps from those contained in the merged data set. In this extra data set, each entry represents a bed request with the following time stamps:

Bed-request time: the time when a bed request is submitted to the bed management unit (BMU);

Bed Allocation: the time when a bed is allocated for the requesting patient;

Bed Confirmation: the allocated bed is confirmed (e.g., by ED nurses if the requesting patient is an ED-GW patient);

Request Completion: the bed-request is completed and the requesting patient is admitted to the allocated bed.

This extra data set has to be extracted from an external IT system which is different from NUH's own system, i.e., it is obtained through a different source from where we get the original four data sets. Due to resource constraints, we have only obtained data from June 1 to December 31, 2008, and June 1 to December 31, 2010 (14 months in total).

By matching patient ID and case number, we link this 14-month data set with the merged data set. Thus, for an ED-GW patient whose bed-request is submitted within the 14 months, we know the confirmation and completion time of the bed-request. Given this information, we can estimate the empirical distributions of allocation delays (Section 4.1 of [25] and Section 9). In addition, the new time stamps in this extra data set allow us to estimate the bed-request distributions for ICU-GW and SDA patients.

3 Changes in discharge distribution and operating environment

From July 2009 to December 2009, NUH started a campaign to discharge more patients before noon. This early discharge campaign gathered momentum and by December 2009, a new and stable discharge distribution emerged. In Section 3.1, we present the empirical discharge distributions in Period 1 and Period 2. In Section 3.2, we describe the measures that NUH introduced in the second half of 2009 to achieve the new discharge distribution. We also explain the reason for choosing Period 1 and Period 2. In Section 3.3, we discuss the changes in the operating environment between 2008 and 2010 and why they preclude us from using the empirical comparison of performance measures for Periods 1 and 2 to evaluate the early discharge policy.

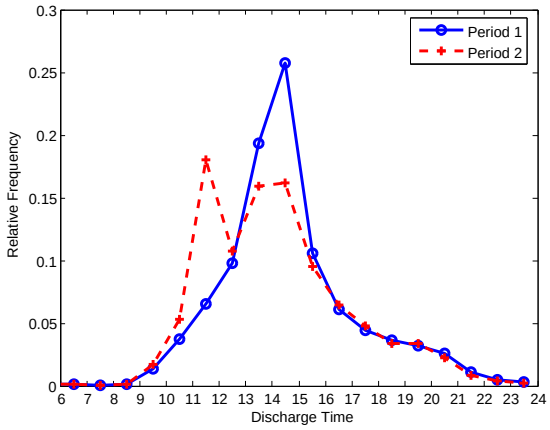


Figure 3: Discharge time distributions: Period 1: from January 1, 2008 to June 30, 2009; Period 2: from January 1, 2010 to December 31, 2010

3.1 Discharge distributions in Periods 1 and 2

Figure 3 plots the hourly discharge distributions in the two periods. For each hour, the corresponding dot in the figure represents the percentage of patients who are discharged from GWs during that hour. Table 1 lists the corresponding numbers for the two discharge distributions. In Period 1, 12.7% of the patients are discharged before noon, and there is a single discharge peak between 2pm and 3pm. In Period 2, 26.1% of the patients are discharged before noon, more than double the percentage in Period 1. It is evident from Figure 3 that there is a new discharge peak between 11am to 12pm in Period 2. In Period 1, as many as 26.3 patients are discharged per hour during the peak time (2-3pm). In Period 2, the peak number of discharge is reduced to 21 patients between 11am and 12pm, and the average number of patients discharged in the original peak hour (2-3pm) is reduced to 18.7 patients. The average discharge hour is moved from 14.6 to 14.1, a half-hour earlier. These statistics indicate that NUH has obtained a satisfactory compliance rate in discharging more patients before noon in Period 2.

Notwithstanding the increase in the proportion of discharges before noon in Period 2, in [25] the authors demonstrate through a high-fidelity stochastic model that the early discharge distribution achieved by NUH as of December 2009 has a limited effect on the waiting time statistics of ED-GW patients (see results in Section 5.2 of that paper). They also show that to achieve a significant improvement in waiting time statistics requires moving the first discharge peak to an earlier time. The next section explains how NUH implemented the early discharge policy.

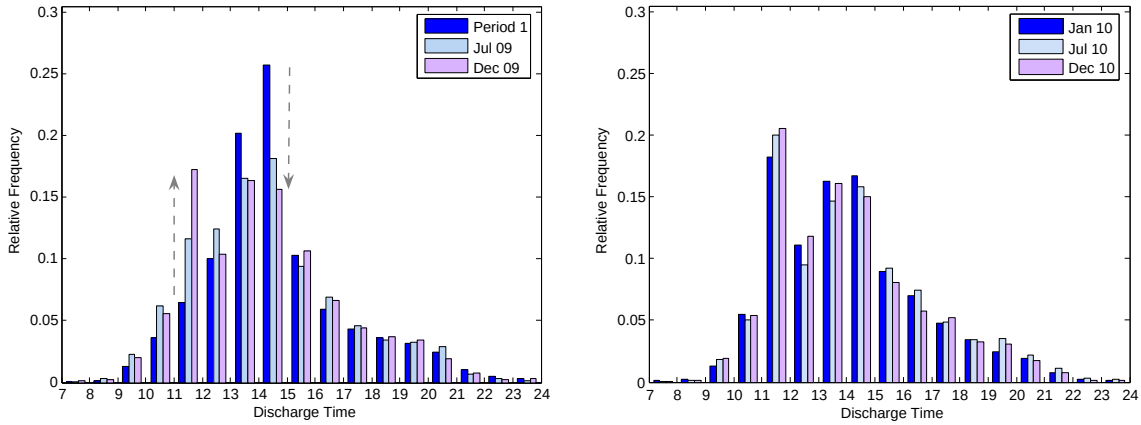
3.2 Implementation of the early discharge policy

The discharge process at NUH is typical of most hospitals [2, 10, 26]. Discharge planning usually begins a day or two prior to the anticipated discharge date. On the day of discharge, the attending physician makes the morning round, confirms the patient's condition, and writes the discharge order. The nurses document the order and prepare the patient for discharge. Finally, pharmacy delivers discharge medication if needed. Obviously, a variety of factors can affect the actual discharge time, such as when the doctor performs the rounds, when pharmacy delivers the medication, and transportation arrangements to send the patient home or to step-down facilities.

To expedite the discharge process and have more patients discharge before noon, NUH began an early discharge campaign from July 2009. The campaign initially started with a small number

Dis. time	Period 1	Period 2
0-1	0.15%	0.12%
1-2	0.15%	0.15%
2-3	0.11%	0.11%
3-4	0.09%	0.11%
4-5	0.10%	0.08%
5-6	0.11%	0.11%
6-7	0.15%	0.12%
7-8	0.07%	0.08%
8-9	0.16%	0.16%
9-10	1.32%	1.68%
10-11	3.69%	5.35%
11-12	6.55%	17.99%
12-13	9.77%	10.75%
13-14	19.39%	15.91%
14-15	25.74%	16.17%
15-16	10.56%	9.49%
16-17	6.08%	6.49%
17-18	4.46%	4.74%
18-19	3.68%	3.36%
19-20	3.24%	3.34%
20-21	2.55%	2.22%
21-22	1.06%	0.85%
22-23	0.47%	0.37%
23-24	0.32%	0.22%

Table 1: Discharge time distributions from general wards: Period 1: January 1, 2008 to June 30, 2009; Period 2: January 1, 2010 to December 31, 2010.



(a) Distributions in Period 1, and July and December of 2009

(b) Distributions in Jan, July, and Dec of 2010

Figure 4: Discharge time distributions during and after the implementation of early discharge

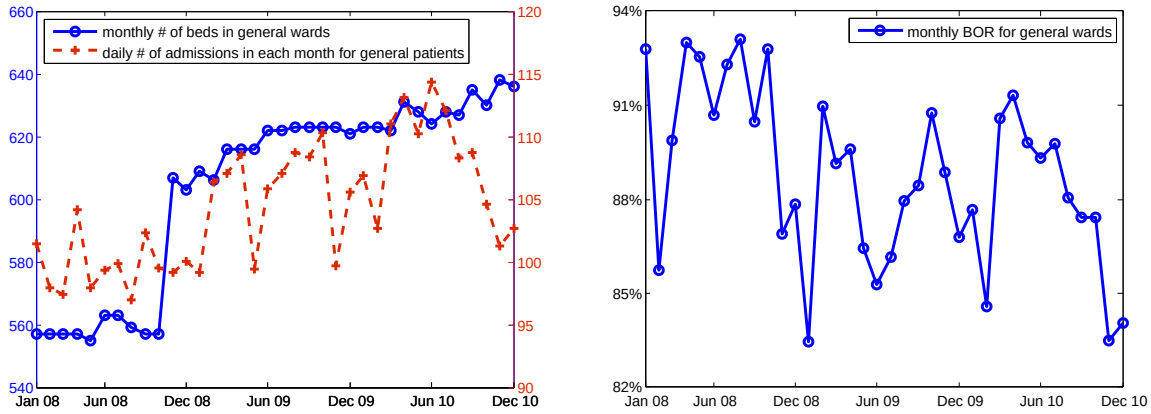
of wards, and was later expanded to the entire inpatient department. By December 2009, the early discharge was completely in effect. Hospital managers have worked closely with physicians, nurses, and patients to promote the campaign. Some of the initiatives include:

Physicians: Physicians in some specialties do two rounds per day (instead one morning round). They try to finish the first round before 10 am, so that some patients can leave before 12 noon. The second round is at about 2-3pm, and more patients can be discharged in the late afternoon.

Discharge lounge: New discharge lounges are added to several wards. Patients waiting for medicines or transportation can wait in the lounge to free up hospital beds.

Day-minus-1-discharge: Physician and nurses identify discharge needs as early as possible and prioritize tests (or other clearance) accordingly. Nurses begin to prepare discharge documents and medicine before the day of discharge.

The early discharge policy is not only costly to implement, but also requires time to attain a high rate of compliance. Indeed, we observe a "stabilizing" process in the discharge patterns when the new policy was being implemented in NUH. Figure 4a compares the Period 1 discharge distribution with the distributions for July and December 2009. As early as May 2009, the peak discharge value decreases from 25.7% to 20.0% comparing to other months in Period 1, while more patients are discharged between 11am and noon. However, at that time there is no explicit second peak in the discharge distribution. From May through September, the value of the original peak (between 2 and 3pm) keeps decreasing, and the proportion of patients discharged between 11am and noon keeps increasing. Till September 2009, a new peak between 11am and noon with a peak value of 14.4% emerge. In December 2009, the new peak is even higher (peak value 17.3%) than the 2-3pm peak value (16.4%). Figure 4b compares the discharge distributions in some selected months of 2010. We can see the distribution stabilizes in 2010. The above observations explain why we choose Periods 1 and 2, since they correspond to before and after implementation of the early discharge policy. We exclude July through December 2009 to avoid potential bias resulting from discharge distribution instability.



(a) Monthly admission rate and number of general beds

(b) Monthly BOR

Figure 5: The monthly admission rate from general patients, the monthly average number of general beds, and monthly BOR from January 2008 to December 2010.

As introduced in the main paper, early discharge policy has been recommended by many previous studies [4, 32] and government agencies [8]. However, few hospitals have reported to implement the policy with any success. For example, studies mention "limited success in achieving discharges by noon" in certain hospitals [28, 32], or that the policy was only experimented in a few wards [8]. Several hospitals claim that they have implemented or tried to implement the early discharge policy [8, 14, 23, 24, 31, 32], but its impact on hospital performances has not been well documented. To our best knowledge, NUH is one of the few hospitals that have successfully implemented the early discharge policy in the *entire* hospital and achieved satisfactory compliance rate as of December 2009. High-fidelity data is also available for us to compare the performances empirically before and after the implementation. Sections 4 and 5.2 below show empirical results on comparing the waiting time statistics and overflow proportions between the two periods, respectively.

3.3 The changing operating environment

The total numbers of admissions to GWs are 36473 in 2008, 38509 in 2009, and 39429 in 2010. Figure 5a plots the monthly admission rate (the red curve) from January 2008 to December 2010. To meet the increasing demand, NUH has increased general bed capacity over the three years. The blue curve in Figure 5a plots the monthly average number of beds in GWs. As a result, we observe a change in the bed occupancy rate (BOR). BOR is a key performance measure which reflects the utilization of beds in a specified period (see the end of this section for a rigorous definition). Figure 5b plots the monthly BORs of the GWs from 2008 to 2010. The average BOR is 90.3% in Period 1, and 87.6% in Period 2. Therefore, Period 2 has a 2.7% reduction of BOR.

Sections 4 and 5.2 will empirically show that both the waiting time statistics and overflow proportion are reduced in Period 2, indicating an improvement in the hospital operations. We note, however, such pure empirical comparisons between the two periods cannot determine the effectiveness of the early discharge policy, due to changes in the operating environment in both periods. From queueing theory, we know that reduced utilization (i.e., BOR in hospital settings) could lead to a reduction in waiting time. Correspondingly, the overflow proportion can be reduced since most overflows are trigged to avoid excessive long waiting. So can the performance improvements in

Period 2 be due to the BOR reduction? To evaluate the effect of early discharge and potentially other operational policies, a model is needed, which is the main focus of [25].

Definition for BOR: BOR is always defined for a specific group of beds. The group can be all beds in a ward or all beds in all general wards. In this document, our default group is all beds in all general wards when no group is specified. For a given group of beds and a given period, BOR is defined as (see pages 10-11 of [21]):

$$\text{BOR} = \frac{\text{Total Inpatient Days of Care}}{\text{Total Bed Days Available}} \times 100, \tag{1}$$

where the *Total Inpatient Days of Care* equals the sum of patient days among all patients who have used a bed in the group within the same period, and patient days of a patient equals the number of days within the period that a patient occupies any bed in the group. Note that patient days of a patient is almost equal to the LOS (see Section 7), except that the patient day of same-day discharge patients is 1, while LOS uses 0 for same-day discharge patients; also LOS may include days outside the given time period. *Total Bed Days Available* is equal to the sum of bed days available among all beds in the group, where bed days available of a bed is the number of days within the time period that bed is available to be used for patients.

4 Waiting time for ED-GW patients

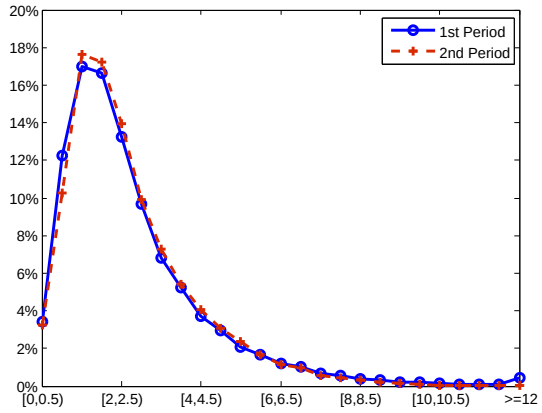
As introduced in Section 2.2, ED-GW patients are those patients who have completed treatment in the emergency department and need to be admitted into a general ward. In this document, we define the *waiting time* of an ED-GW patient as the period between bed-request and exit from the ED, which is consistent with the performance measure that the Singaporean Ministry of Health (MOH) uses to monitor all its public hospitals (e.g., see [27]). This waiting time is also known as the "ED boarding time", a commonly used term in US health systems [17, 30].

Note that the waiting time defined here slightly differs from the one defined in [25], which uses patient's admission time as the endpoint instead of the exit time from the ED (e.g., see Figure 1 and 5 in that paper). Therefore, the waiting time reported in this document is more conservative, because it excludes delays during transportation to the GW and admission to the bed. From the analysis in Section 4.1 of [25] and Section 9, one can see that the gap between the time leaving ED and the actual admission time is a part of the post-allocation delay.

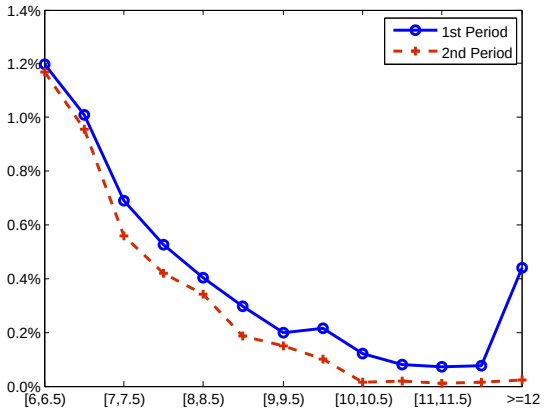
The main purpose of using admission time as the endpoint in [25] is for model calibration. The proposed stochastic model in that paper assumes a patient's service begins after admission, so the duration between arrival (bed-request) and admission becomes the waiting time. To calibrate the model output with empirical performances (e.g., see Figure 7 and Table 4 of [25]), one needs to keep consistent in calculating the waiting times. Whereas in this document, our main focus is the empirical study. Thus, we report the waiting time statistics in the same way as the MOH's definition. Nevertheless, noting that the gap between ED exit and admission to a ward (bed-request completion) is about 18 minutes on average in each Period, it is short compared to the overall average waiting time. Thus, the basic trends of the waiting time statistics (hourly and specialty-level performances) are similar under the two definitions, and most of our observations hold no matter which definition is used.

4.1 Distribution of waiting time

Figure 6a shows the empirical distributions of waiting times for all ED-GW patients in Periods 1 and 2. The bin size is 0.5 hour, and points falling beyond 12 hours are lumped together into the



(a) Distribution between 0 and 12 hours



(b) Distribution between 6 and 12 hours

Figure 6: Empirical distributions of the waiting times for ED-GW patients. The bin size is 0.5 hour, and points falling beyond 12 hours are lumped together into the last bin.

last bin. For hospital management purpose, we are particularly interested in those patients with excessive long waiting times. Thus in Figure 6b, we provide a detailed plot for the waiting time distribution between 6 and 12 hours. The shapes of the overall distribution curves look similar in both periods. The tail distributions, however, exhibit significant differences.

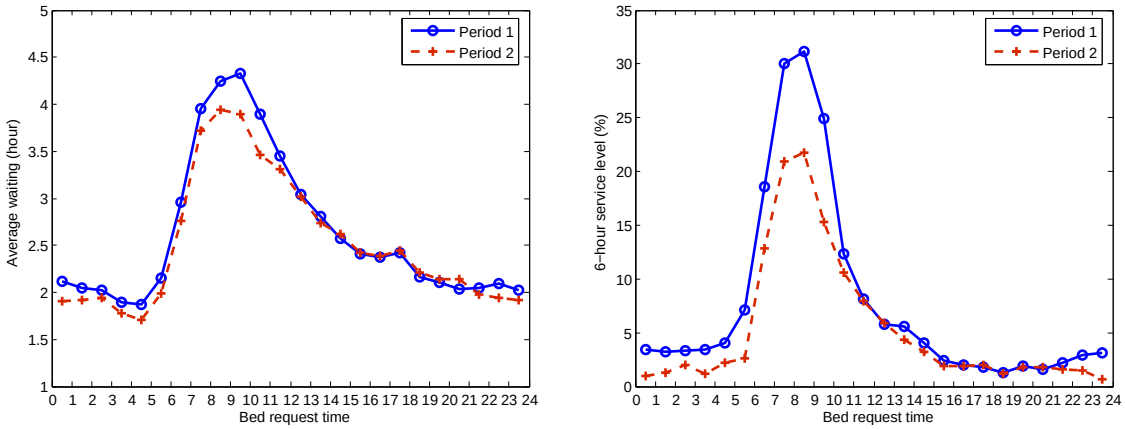
In the next three subsections, we first present findings on the average waiting time in Section 4.2. Then in Section 4.3, we focus on the tail distributions of waiting time, i.e. service levels. Finally in Section 4.4, we compare the waiting time statistics (both the average and tail) for the nine specialties.

4.2 Average waiting time

As reported in Table 2, the average waiting time for all ED-GW patients is 2.52 hours for Period 1, and 2.46 hours for Period 2, a reduction of 3.6 minutes. Thus, there is no significant difference between the two periods.

The waiting time of an ED-GW patient depends heavily on her bed-request time. Figure 7a shows the hourly average waiting time, which is calculated by averaging the waiting times of patients requesting beds in each hour. Figure 7a is similar to Figure 1(a) in [25] except that we calculate the waiting times differently. Table 3 lists the corresponding numerical values for Figure 7a. The figures and table all show a time-dependent feature of the average waiting time. Patients requesting beds in the morning, between 6am and 12noon, experience much longer average waiting times than patients requesting beds in other hours. Comparing the two periods, we see certain reductions in Period 2 among those patients with long average waiting times (requesting beds in the morning). For patients requesting beds from 12 noon to midnight, however, their average waiting times do not show much difference.

We note that the average waiting time in each hourly interval is above 1.5 hours in both periods. Figure 6a also shows that only around 4% of patients wait less than 0.5 hour. This indicates that the majority of ED-GW patients have to wait for a certain amount of time, no matter when they request beds or whether beds are available. This observation relates to our findings on the allocation delays that will be discussed in Section 9, which is also a key element in building the model (see Section 4.1 of [25]).



(a) Average waiting times

(b) 6-hour service levels

Figure 7: Waiting times statistics for ED-GW patients by bed request hour

	Period 1	Period 2
bed request number	35452	25285
average waiting time (hour)	2.52	2.46
f(W $>$ 4)	15.73%	15.15%
f(W $>$ 6)	5.34%	3.97%
f(W $>$ 8)	1.90%	0.86%
f(W $>$ 10)	0.79%	0.09%

Table 2: The average waiting time and x-hour service levels for ED-GW patients.

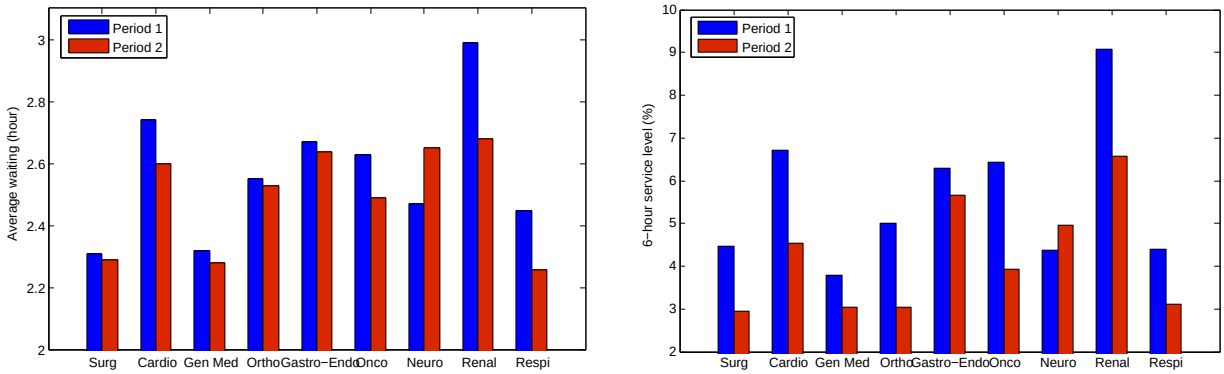
4.3 Service levels

In addition to average waiting times, we focus on *x-hour service level*, denoted by $f(W > x)$, that is defined as the fraction of ED-GW patients who have to wait x hour or longer before exiting the ED. Here W denotes the waiting time of a typical ED-GW patient. Table 2 reports the overall 4-hour, 6-hour, 8-hour and 10-hour service levels in the two periods. The 6-, 8-, and 10-hour service levels among these two periods are significantly different. The 6-hour service level decreases from 5.34% in Period 1 to 3.97% in Period 2, a 26% reduction. The 10-hour service level changes even more dramatically, from 0.79% to 0.09%, an 89% reduction. Note that the total number of bed requests (see the first row) differs significantly in the two periods. This is because Period 1 contains 18 months whereas Period 2 contains 12 months. The average monthly number of bed requests is 1970 and 2107 for Period 1 and 2, respectively.

Figure 7b, which shows the 6-hour service level with respect to bed-request hour, is similar to Figure 1(b) in [25], except that we calculate the waiting times differently. Table 3 lists the corresponding numerical values. We also observe a time-dependent feature of the 6-hour service level. Patients requesting beds between 6am and 12noon have a much higher chance of waiting more than 6 hours than patients requesting beds in other hours. In Period 1, about 1 out of 3 patients requesting beds between 8 and 9am have to wait more than 6 hours. Comparing the two periods, the peak value of the 6-hour service level (8-9am) decreases from 31% to to 22% in Period 2. Table 3 also lists the 4-, 8-, and 10-hour service levels with respect to the bed-request hour. The 8-hour and 10-hour service levels are greatly reduced in each hour in Period 2.

	per	bed-request hour											
		1	2	3	4	5	6	7	8	9	10	11	12
req. dist.	1	4.33	3.48	2.71	2.40	1.81	1.71	1.68	1.50	1.62	2.41	3.51	4.62
(%)	2	4.31	3.33	2.57	2.09	1.65	1.67	1.53	1.67	1.77	2.57	3.41	4.98
avg. wait	1	2.12	2.05	2.02	1.89	1.87	2.15	2.96	3.95	4.24	4.33	3.89	3.45
(h)	2	1.91	1.92	1.94	1.78	1.71	1.99	2.76	3.72	3.94	3.89	3.47	3.31
f(W > 4)	1	8.34	6.24	5.93	5.64	6.22	10.87	23.49	42.59	52.70	56.84	48.79	32.82
(%)	2	6.88	5.95	5.24	3.02	3.37	7.57	23.20	43.94	49.22	44.68	36.54	30.84
f(W > 6)	1	3.39	3.24	3.33	3.41	4.04	7.08	18.62	30.02	31.13	24.91	12.30	8.11
(%)	2	1.01	1.31	2.00	1.13	2.16	2.60	12.89	20.90	21.70	15.25	10.56	7.95
f(W > 8)	1	2.80	2.43	2.81	2.59	2.80	4.45	11.24	9.57	7.65	4.21	2.89	2.20
(%)	2	0.18	0.83	0.92	0.76	0.96	1.18	3.09	5.94	3.36	2.16	2.55	1.91
f(W > 10)	1	2.35	2.27	1.98	1.18	1.87	1.15	2.35	1.50	1.22	0.82	0.56	0.49
(%)	2	0.00	0.12	0.31	0.57	0.24	0.00	0.00	0.24	0.00	0.15	0.12	0.00
	per	bed-request hour											
		13	14	15	16	17	18	19	20	21	22	23	24
req. dist.	1	5.46	6.18	6.30	6.61	6.14	6.16	5.64	4.86	5.07	5.55	5.12	5.12
(%)	2	5.73	5.99	6.58	6.84	5.92	5.92	5.40	5.09	5.20	5.62	5.20	4.98
avg. wait	1	3.04	2.81	2.58	2.41	2.37	2.42	2.17	2.10	2.04	2.05	2.09	2.03
(h)	2	3.02	2.73	2.62	2.42	2.39	2.44	2.21	2.14	2.14	1.98	1.94	1.92
f(W > 4)	1	22.99	20.40	16.88	14.94	12.37	10.77	7.20	6.85	6.40	5.95	6.94	6.44
(%)	2	25.45	20.59	17.43	12.83	11.03	11.49	8.57	8.16	8.14	6.41	7.15	6.20
f(W > 6)	1	5.79	5.61	4.03	2.39	1.98	1.79	1.30	1.86	1.56	2.24	2.92	3.08
(%)	2	5.93	4.36	3.19	1.91	1.87	2.00	1.17	1.79	1.83	1.62	1.45	0.71
f(W > 8)	1	1.29	1.46	0.58	0.30	0.60	0.50	0.35	0.75	0.78	1.32	2.09	2.09
(%)	2	1.10	0.33	0.60	0.17	0.00	0.73	0.22	0.78	0.46	0.35	0.23	0.48
f(W > 10)	1	0.05	0.23	0.04	0.17	0.05	0.23	0.20	0.46	0.56	1.02	1.82	1.38
(%)	2	0.14	0.00	0.06	0.00	0.00	0.27	0.00	0.08	0.08	0.00	0.15	0.08

Table 3: The average waiting times and service levels by bed-request hour.



(a) Average waiting times

(b) 6-hour service level

Figure 8: Waiting times statistics for each specialty.

4.4 Waiting time statistics for each specialty

Figures 8a and 8b plot the average waiting times and 6-hour service levels for the nine specialties in the two periods. Table 4 shows the corresponding numerical values, and contains statistics for other service levels.

We make two observations. First, the nine specialties exhibit similar average waiting time and 6-hour service levels in each period (especially in Period 2). This balanced result could have been achieved through years of continual adjustment in resource allocation (e.g., bed and ward allocation) and a proper overflow policy (see Section 5.2). Renal and Cardiology patients show longer average waiting times than the overall average, while Surgical, General Medicine, and Respiratory patients show shorter average waiting times than the overall average. The potential reasons could be that (i) Surgery, General Medicine, and Respiratory wards have relatively low BORs (see Table 5); moreover, patients from Surgery and General Medicine can be overflowed to wards of other specialties easily since they have less specialized requirements; (ii) Renal and Cardiology wards have high BORs, and these patients need more specialized care and equipment (e.g., dialysis for Renal patients, telemetry beds for Cardiology patients) so it is more difficult for them to be overflowed.

Second, comparing the two periods, we can see that the average waiting time does not change much for each specialty except for Renal and Respiratory. Meanwhile, the 6-hour service level exhibits a significant reduction in Period 2 for each specialty except Neurology. These observations are consistent with what we observed from the hospital-level statistics (see Table 2). They all suggest that patients with long waiting times (as noted, a very small amount) benefit more in Period 2 than most patients.

5 Wards

In this section, we report the statistics on the 19 GWs included in our study. As mentioned, we call the beds in these wards "general beds".

In NUH, each GW contains a number of beds in close proximity. The wards are relatively independent of each other, with each having its dedicated nurses, cleaning team and other staff members. There are usually multiple rooms in each ward. A room is equipped with 1 to 8 beds, depending on the ward "class", and is shared by patients of the same gender. In general, class C wards have 8 beds per room, class B wards 4 or 6 beds per room, and class A wards 1 or 2 beds

	Period	Sample size	avg. wait (h)	f(W > 4) (%)	f(W > 6) (%)	f(W > 8) (%)	f(W > 10) (%)
Surg	1	6078	2.31	13.29	4.46	1.40	0.61
	2	3926	2.29	12.66	2.95	0.76	0.05
Card	1	5437	2.74	19.22	6.71	2.21	0.74
	2	4011	2.60	18.08	4.54	0.85	0.02
Gen Med	1	7913	2.32	12.26	3.80	1.44	0.80
	2	6176	2.28	11.92	3.04	0.65	0.08
Ortho	1	4557	2.55	15.16	5.00	1.73	0.75
	2	2899	2.53	13.42	3.04	0.83	0.07
Gastro-Endo	1	3348	2.67	18.43	6.30	2.42	0.96
	2	2309	2.64	19.23	5.67	1.08	0.09
Onco	1	1586	2.63	17.40	6.43	3.22	1.32
	2	1271	2.49	16.76	3.93	0.79	0.08
Neuro	1	2669	2.47	15.36	4.38	1.57	0.52
	2	1979	2.65	18.85	4.95	0.76	0.10
Renal	1	2315	2.99	23.24	9.07	3.59	1.43
	2	1686	2.68	19.51	6.58	2.02	0.42
Respi	1	1549	2.45	14.33	4.39	1.23	0.39
	2	1028	2.26	12.16	3.11	0.58	0.00

Table 4: Waiting time statistics for ED-GW patients from each specialty.

per room (see details in [19]). Stays in class B2 or C wards are eligible for heavy subsidy from the government, thus the daily expenses in these "subsidized wards" are much less than the expenses in class A or B1 wards. As a result, there is a greater demand for the subsidized wards.

Physicians always prefer to have their patients stay in the same wards to save rounding time. Hospital can also achieve a better match between patient needs and nurse competencies by doing so. Therefore, NUH designates each general ward to serve patients from only one or two (rarely three) specialties. We call the ward's designated specialty its *primary* specialty. Around September to December 2008, NUH changed the primary specialties for several wards to better match the demand and supply of bed for each specialty, a reaction to the big capacity increase in late 2008 (see Figure 5a). Since most of our reported statistics in this section relate to the ward primary specialties, we exclude the period before the re-designated specialties became operational for consistency. The term "reduced Period 1", therefore, refers to the remaining time in Period 1 after the re-designated specialties took effect. The start time for the reduced Period 1 for each affected ward depends on the time of specialty re-designation; the end time is fixed at June 30, 2009. Thus, the duration of the reduced Period 1 may differ for each ward, since the re-designated specialty could take effect at different times.

For wards with no changes in their primary specialties, we use data points from the entire Period 1 to calculate the statistics; otherwise, we use the reduced Period 1. We calculate statistics for Period 2 using the data points of the entire period, because no speciality re-designation occurred. Table 5 lists the start month of (reduced) Period 1 for each ward. For example, Ward 52 was re-designated as an Orthopedic ward from November 2008, and Ward 54 a Surgery/Orthopedic ward from March 2009.

5.1 Capacity and BOR

Figure 5b plots the monthly BOR for all general wards from January 2008 to December 2010. See Section 3.3 for the definition of BOR. Figure 5b indicates that the monthly BOR fluctuates between 80% and 95%. The average BOR is 90.3% for Period 1 and 87.6% for Period 2. In fact, if we exclude January to October 2008, the average BOR for the remaining Period 1 is about 87.4%, which is similar to Period 2. This suggests that NUH has successfully increased its bed capacity, resulting in BOR stabilization despite significant increases in patient admissions from January 2008 to December 2010. The total number of general beds increased from 555 beds as of January 1, 2008 to 638 beds as of December 31, 2010.

Not surprisingly, BOR is ward dependent. Table 5 lists the number of beds in each ward (as of January 1, 2008 and December 31, 2010, respectively), the primary specialties, and the BORs in Periods 1 and 2. The BORs for all 19 wards are also plotted in Figure 11. We make the following observations: (i) BORs of dedicated wards (43, 56, 57, 58, 63, 64) are generally high, most exceeding 90%, with the exceptions of Orthopedic wards 51 and 52 which have much lower BORs for both periods; (ii) class A/B1 wards (66, 76, 78, 86) have lower BORs than other wards because they are not government-subsidized; (iii) ward 44 has a much lower BOR than other Medicine wards, mainly because half of its capacity serves infectious respiratory patients who cannot share rooms with other patients; and (iv) comparing the BORs for the two periods shows no consistent pattern of increase or decrease.

We note two other facts. First, we cannot calculate BORs for each specialty, because some of them share wards and patients can be overflowed to non-primary wards. The BOR for each ward reflects which specialty generally has a higher BOR though. Second, BOR uses inpatient day in the calculation which only takes integer values (see Section 3.3), so it is different from utilization rate which uses service time. We report the BORs instead of the ward utilization rates in this document, while the main paper reports the latter for model calibration purposes (see Section 5.1 of [25]). From our calculation, the BOR is slightly higher than the corresponding utilization for most wards, but the two values are very close, typically differing by only 1% to 2%.

5.2 Overflow proportion and BOR share

Overflow proportion

This document and the main paper [25] define overflow proportion as the number of "overflow" admissions (i.e. patients admitted to non-primary wards) divided by the total number of admissions. Note that the number of admissions here (for both the numerator and denominator) include the initial admissions and re-admissions to the GWs. For example, suppose a patient is initially admitted to GW, then transfers to ICU, and finally transfers back to GW. For this patient, we count the initial admission and the second transfer (ICU to GW) in the calculation, and consider the transfer as a re-admission to the GW. Similarly, we count a patient transferring from one GW to another GW as a re-admission to the receiving GW. Basically, no matter whether it is an initial admission or a transfer event, as long as the receiving ward is a general ward, we consider this event as one admission in the calculation of overflow proportion. The reason is that BMU needs to find a receiving ward when such an event occurs, so there is a chance that the patient will be assigned to a non-primary ward.

At NUH, the overall overflow proportion is 26.95% and 24.99% for Periods 1 and 2, respectively. Next, we present empirical results of overflow proportions at the *ward level* and *specialty level*. The overflow proportion for a ward is defined as the number of overflow admissions to this ward divided by the total number of admissions to this ward. The overflow proportion for a specialty is defined as the number of overflow admissions from this specialty divided by the total number of admissions

Ward	Prim. specialty	# of beds		Per 1 start	BOR (%)	
		Jan 08	Dec 10		Per 1	Per 2
41	Surg, Card	44	44	Feb 09	90.9	92.0
42	GM, Respi	33	44	Nov 08	86.4	92.2
43	Surg	44	44	Jan 08	93.4	88.9
44	Respi, Surg	14	44	Mar 09	79.0	80.3
51	Ortho	39	39	Jan 08	76.7	67.5
52	Ortho	22	26	Nov 08	74.4	75.3
53	GM, Neuro	46	46	Jan 08	96.8	97.1
54	Surg, Ortho	50	50	Mar 09	80.7	77.6
55	Renal	44	33	Jan 08	91.7	86.5
56	Card	17	17	Nov 08	90.1	95.1
57	Neuro	14	14	Jan 08	97.3	96.5
57O	Onco	24	24	Jan 08	93.9	93.2
58	Onco	24	24	Jan 08	90.2	91.7
63	Card	43	44	Jan 08	95.5	96.1
64	Gastro	46	50	Jan 08	94.2	92.8
66	Med, Surg	31	34	Feb 09	86.9	86.8
76	Med, Card	18	18	Jan 08	90.0	94.6
78	Onco, Surg, Ortho	25	25	Mar 09	83.0	82.8
86	Onco	8	14	Jan 08	89.6	87.5
Total General Beds		555	638		90.3	87.6

Table 5: Designated (primary) specialties and BORs for the 19 general wards.

from this specialty. As opposed to the overflow admission, we call an admission a *right-siting* if the patient is admitted to a primary ward. The right-siting proportion is defined in a similar way as overflow proportion, and their sum is equal to 1.

Table 5 lists the primary specialties for the 19 GWs. We call a GW a dedicated ward if there is only one primary specialty; otherwise, we call it a shared ward. Figure 9 compares the overflow proportions for the GWs by period. Table 6 lists the corresponding numerical values. Dedicated wards generally have a lower overflow proportion than the shared wards. Comparing the two periods, most of the wards show reduced overflow proportions in Period 2, with some showing significant reductions (mostly dedicated wards); some wards show a small increase. The only exceptions are ward 44 and 52, which show significant increases in the overflow proportions.

Moreover, comparing the BOR (Table 5) and the overflow proportion (Table 6) for each ward, we can see it is generally true that if the ward has a lower BOR, its overflow proportion will be higher; examples are wards 51, 52, and 54. The only exception is ward 44, which has a low BOR and a low overflow proportion at the same time. In practice, the BMU prefers to overflow class A/B1 patients to a non-primary class A/B1 ward instead of downgrading them to a lower-class primary ward. This also explains why class A/B1 wards have higher overflow proportions than most class B2/C wards, since class A/B1 wards are "pooled" together more often.

Figure 10 compares the overflow proportion for each specialty in Periods 1 and 2. Note that (i) Cardiology, General Medicine, and Neurology patients have significant higher overflow proportions than other specialties, which suggests that these specialties may not have enough beds allocated to them; (ii) the overflow proportions of Surgery, General Medicine, Respiratory, and Orthopedic show significant reductions in Period 2, whereas Gastro-Endo and Neurology show a big increase in

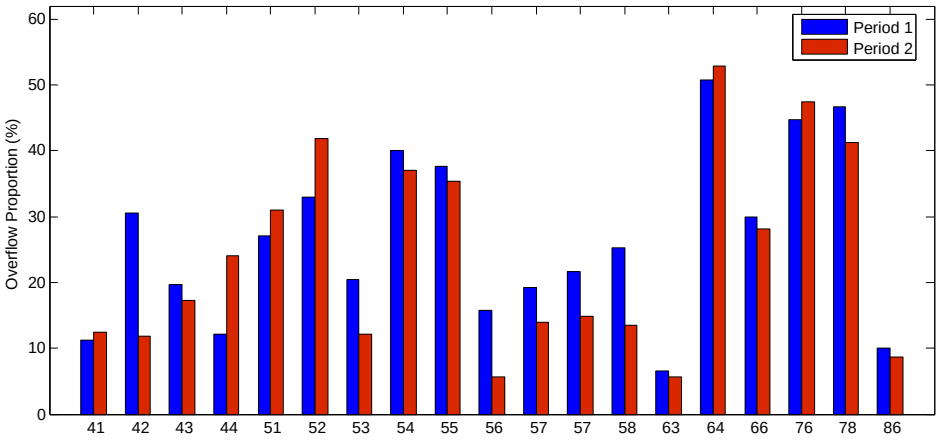


Figure 9: Overflow proportion for each ward in Periods 1 and 2.

Ward	OvFlow proportion (%)		OvFlow BOR share (%)	
	per 1	per 2	per 1	per 2
41	11.2	12.5	11.4	8.3
42	30.5	11.8	23.7	8.9
43	19.7	17.3	25.1	15.5
44	12.1	24.0	14.9	23.2
51	27.1	31.0	14.3	19.4
52	33.0	41.9	21.3	28.5
53	20.4	12.1	15.2	16.6
54	40.1	37.0	25.8	21.6
55	37.7	35.3	29.0	25.1
56	15.7	5.7	12.6	3.2
57	19.2	14.0	13.9	11.6
57	21.6	14.9	11.8	8.1
58	25.3	13.5	13.5	6.6
63	6.5	5.6	5.2	4.5
64	50.7	52.9	47.0	49.3
66	30.0	28.2	27.9	29.3
76	44.8	47.5	48.4	50.2
78	46.7	41.2	43.0	37.1
86	10.0	8.7	4.0	3.2
Total	27.0	25.0	21.4	19.2

Table 6: Overflow proportion and BOR share for each ward.

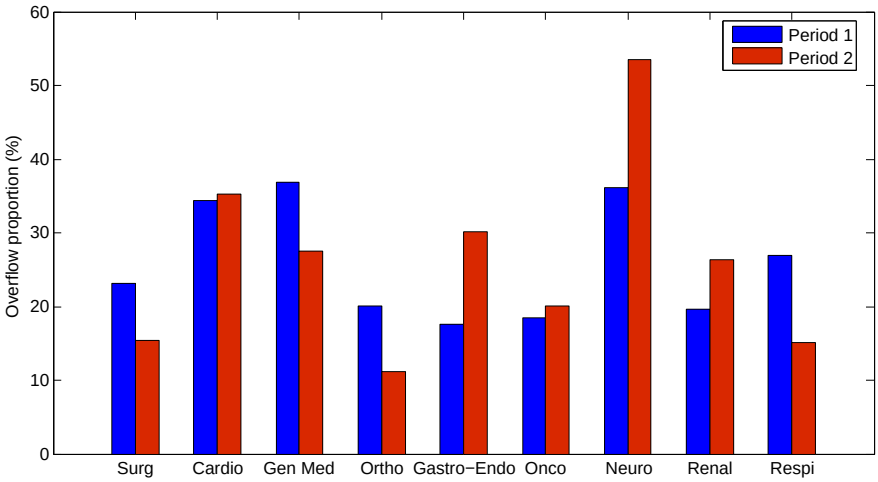


Figure 10: Overflow proportion for each specialty in Period 1 and 2

Period 2.

BOR share

We define *BOR share* of a specialty (or group of specialties) as the BOR of the speciality, or group, divided by the total BOR for a certain ward. To calculate the BOR of one specialty for a given ward, the numerator in Equation (1) counts the total patient days for patients from that specialty who used beds in the ward. The denominator counts the total bed days available for all beds from that ward. Thus, the sum of the BORs from each specialty equals the total BOR of that ward (as reported in Table 5). Correspondingly, the BOR share from each specialty adds up to 1 for that ward.

When modeling beds in a ward as servers, the BOR share resembles the workload share in queueing systems, i.e., out of all "busy" periods, the average proportion of time that these beds are "working" on patients from a particular specialty. The BOR share provides us with a deeper insight into the overflow issue. Patients who are initially assigned to a wrong ward may be transferred to the right ward later (see more discussions in Section 10.2 on such transfers). Typically, this happens a day or two after the patient's initial admission; otherwise, the hospital usually allows the patient to remain in the wrong ward until discharge. The overflow proportion only takes patient count into consideration, without differentiating an overflow patient with a long LOS from an overflow patient with a short LOS, where the latter is always preferred for the right-siting of care. Therefore, we study this BOR share statistic, since it takes patient's LOS into consideration from the BOR calculation.

Figure 11 plots the BORs from primary and non-primary specialties for each ward in Periods 1 and 2. We also refer the two BORs as right-siting BOR and overflow BOR, respectively. Each bar in the figure represents the total BOR for each ward in the corresponding period. Even though the figure does not directly plot the BOR share (since the BOR share from primary and non-primary specialties should add up to 1 for a ward), it gives us some insight regarding the time the ward serves right-siting patients and overflow patients, as well as its "idle" time, when it is not serving patients. Table 6 contains the numerical values for the overflow BOR share for each ward in Periods 1 and 2. Using 1 minus the overflow BOR share obtains the right-siting BOR share. We observe similar features regarding the overflow BOR share and overflow proportions. For example, dedicated

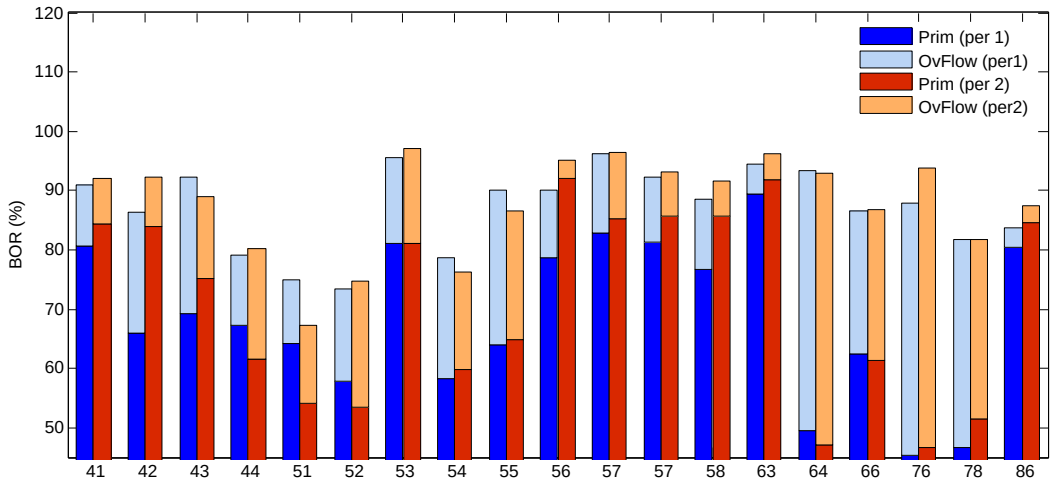


Figure 11: BOR from primary and non-primary specialties for each ward in Period 1 and 2. Each bar height represents the total BOR for each ward in the corresponding period. The y-axis starts from 45%.

wards have lower BOR share from overflow patients, and Orthopedic wards and class A/B1 wards expend more time treating overflow patients. Moreover, most wards in Period 2 show a reduction in overflow BOR share.

Comparing the overflow BOR share with overflow proportion in Table 6, we can see that the overflow BOR share is generally smaller than the corresponding overflow proportion, e.g., ward 54 and 55. This has two implications. First, some overflow patients only stay in the wrong wards for a day or two before transfer to a right ward. Thus, the lower overflow BOR share value (compared to overflow proportion) reflects NUH's efforts on right-siting. Second, the overflow patients have a shorter average LOS compared to the primary patients, even when they do not transfer to a right ward, e.g., ward 58 is dedicated to serve Oncology patients, and most of its overflow patients are from General Medicine with a shorter average LOS. In fact, this also explains why ward 43 shows a higher overflow BOR share value than overflow proportion in Period 1, since most of its overflow patients are from Orthopedic with a longer average LOS than its primary Surgery patients.

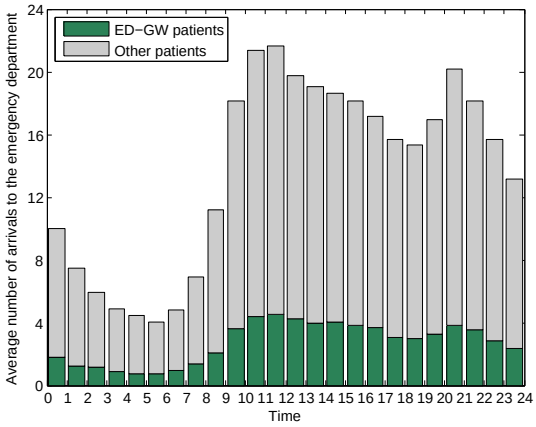
5.3 Shared wards

Excluding class A/B1 wards, NUH has five shared wards, 41, 42, 44, 53, and 54, serve two primary specialties (see Table 5). Each bed in the shared wards is still nominally allocated to a certain specialty, but the nurses in these wards have the flexibility to care for patients from either speciality.

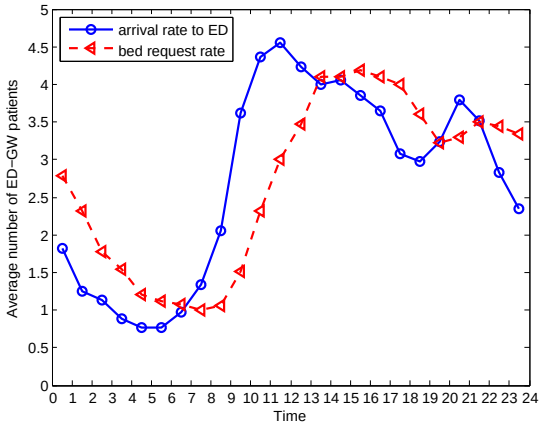
For each of the shared wards and for each period, we calculate the ratio between the BORs of the two primary specialties and the ratio between their admission numbers. We compare these two ratios with the nominal capacity allocation. Table 7 lists these three sets of statistics (in Columns 4-5, 6-7, 8, respectively). First, we can see that the ratios of the BORs and the ratios of admission numbers are close for each ward, except for ward 44 in Period 2 and ward 54 in Period 1. The closeness indicates that the average LOS of the two primary specialties are close. Second, we can see that the ratios in Columns 4-7 are mostly above 80%, and generally exceed the ratios of the nominal bed allocation (last column). This indicates that each ward is still predominantly used by patients from one certain specialty, regardless of the nominal allocation.

Ward	Specialty		Ratio of BOR		Ratio of admissions		Ratio of alloc beds
	Prim. 1	Prim. 2	per 1	per 2	per 1	per 2	
41	Surg	Card	81.45	81.10	81.97	80.27	72.09
42	Gen Med	Respi	94.93	95.66	92.34	93.94	77.27
44	Respi	Surg	72.62	69.26	67.48	59.39	53.33
53	Gen Med	Neuro	86.49	93.50	82.09	89.50	unknown
54	Ortho	Surg	86.53	84.90	73.83	80.31	66.67

Table 7: Shared wards. The ratio of BOR is defined as the BOR from Prim.1 specialty divided by the sum of BORs from its primary specialties (i.e., right-siting BOR). The ratio of admissions or the ratio of allocated beds is defined similarly by just changing BOR to the number of admissions or the number of allocated beds, respectively. The ratios of allocated beds are estimated from the average number of beds in both periods; the nominal bed allocation is unknown for Ward 53.



(a) Arrival rate to emergency department



(b) Arrival rate and bed request rate of ED-GW patients

Figure 12: Hourly arrival rate of all patients to the emergency department, and hourly arrival rate and bed-request rate of ED-GW patients, who will eventually be admitted into general wards. Period 1 data is used.

6 Bed-request from ED-GW patients

Recall from Figure 1 that in both periods, about 64% of the general patients are ED-GW patients. In this section, we focus on the bed-request process from ED-GW patients. In Section 6.1, we study the hourly bed-request pattern of ED-GW patients and show its connection with the arrival process to the emergency department. In the main paper [25], the authors assume a non-homogeneous Poisson process for the bed-request process from ED-GW patients. In Section 6.2 we show some statistical testing results for the non-homogeneous Poisson assumption.

6.1 Bed-request rate

For modeling purposes, the main paper [25] sometimes uses "arrival" and "bed-request" interchangeably; we need to differentiate between them for this online supplement. The arrival time to ED is when a patient shows up at ED from either walk-in or ambulance. The bed-request time is when ED physicians decide to admit a patient after treatment in ED and request a bed for this patient. Only

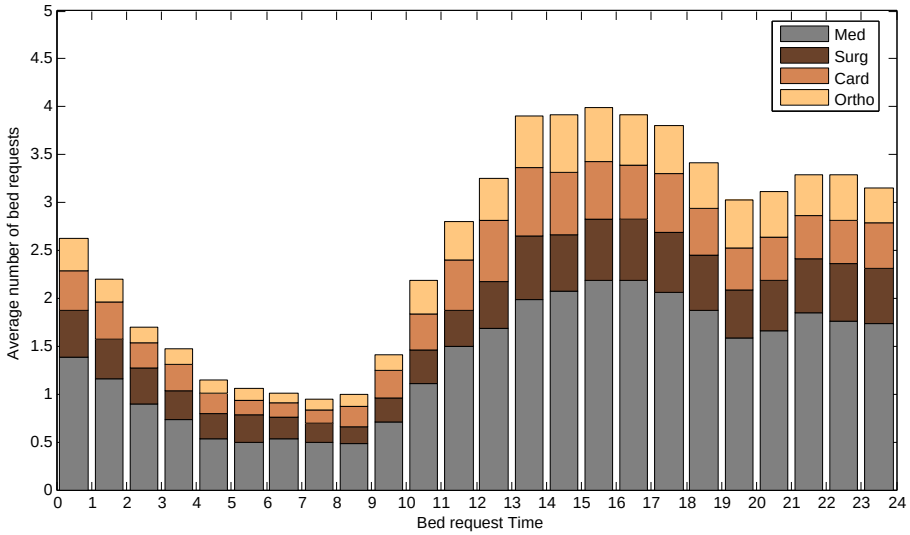


Figure 13: Hourly bed request rate for the major specialties in Period 1. The plot aggregates the five specialties belong to the Medicine cluster and omits Oncology.

about 20% of the arrivals to ED are admitted to the GWs and become ED-GW patients. Figure 12a plots the hourly arrival rate to ED from all patients in Period 1. The green bars represent the arrival rate from those ED-GW patients who will eventually be admitted to a general ward. The grey bars represent the arrival rate from all other patients, who will be directly discharged from the ED or admitted to other wards. The sum of the two arrival rates shows the average number of arrivals to ED from all patients in each hour. Note that the arrival rate (all patients) begins to increase from 7 am, followed by two peaks: a peak between 11am and noon (21.7 per hour) and a peak between 8pm and 9pm (20.2 per hour). This pattern is similar to those observed in hospitals of other countries (e.g., see Figure 1 of [9] and Figure 2 of [33]), indicating that the arrival rate to NUH's emergency department is not unique.

Figure 12a also shows that the proportion of the green and grey bars does not change much throughout the day. About 17% to 22% of patients arriving at the ED become ED-GW patients in each hour, which suggests that the patient mix (ED-GW versus other patients) is quite stable.

Figure 12b demonstrates the connection between ED arrival and bed-request of ED-GW patients. The blue curve shows the arrival rate to ED from ED-GW patients, which is identical to the green bars in Figure 12a. The red curve shows the average number of bed requests from ED-GW patients during each hour. We use the term "hourly bed request rate" to denote the number of beds requested by ED-GW patients in each hour. The bed request rate starts to increase from 7am, and reaches three or more per hour between noon and midnight. The peak is between 1 pm and 5pm (4.2 per hour). If we compare the two curves in Figure 12b, we can see their shapes are similar and the red curve seems to be a horizontal shift of the blue curve. This depicts the relationship between the arrival process to ED and the bed request process of ED-GW patients: when an ED-GW patient arrives at the emergency department, it takes about two hours to receive treatment (plus the possible waiting time) before a physician decides to admit and makes a bed-request.

Figure 13 compares the hourly bed request rate from ED-GW patients by cluster (Medicine, Surgery, Cardiology, and Orthopedic); we aggregate the five specialties belonging to the Medicine cluster (see Section 2.3) and omit Oncology due to its small volume. This figure shows that the proportion of the clusters changes little over time, suggesting that patient distribution is stable in

each hour. It is also consistent with our observation that the bed request rate curves from each specialty have similar shapes (figures not shown here).

Figures 12 and 13 use Period 1 data. Using Period 2 data show similar patterns/phenomena. However, the average arrival and bed request rates both increase in Period 2, since more patients visit the hospital (also see Section 3.3). Therefore, it is not suitable to combine the data of the two periods to show arrival rates to the ED and the bed-request rates.

6.2 Testing the non-homogeneous Poisson assumption

Brown et al. [5] proposed a method to test non-homogeneous Poisson arrival processes. We apply this method to NUH data to test the bed-request process from ED-GW patients. The null hypotheses of our test is that the bed-requests of ED-GW patients form an inhomogeneous Poisson process with piecewise-constant arrival rates.

To perform the test, we follow the procedures described in [5]. First, we divide each day into 7 time blocks: 0am-2am, 2am-4am, 4am-9am, 9am-11am, 11am-13pm, 13pm-18pm, and 18pm-0am. Note that we do not use blocks of equal length. We choose these blocks so that within each of them, the hourly arrival rates are close for the included hours (see the red curve in Figure 12b). We call a block on a certain day a *time interval*, e.g., 2am-4am on May 1, 2008 is a time interval. The blocks we choose also ensure that we have enough data points in each time interval. Second, for each time interval i, we collect the bed-request time stamps belonging to that interval and transform the time stamps in the same way as introduced in [5]. That is, let T_j^i denote the j^{th} ordered bed-request time in the i^{th} interval $[T_{\text{start}}^i, T_{\text{end}}^i)$, $i = 1, \ldots, I$, where I denotes the total number of intervals. Let $J(i)$ denote the total number of bed-requests in the i^{th} interval, and define $T_0^i = T_{\text{start}}^i$ and $T_{J(i)+1}^i = T_{\text{end}}^i$. Then we have $T_{\text{start}}^i = T_0^i \le T_1^i \le \cdots \le T_{J(i)}^i < T_{J(i)+1}^i = T_{\text{end}}^i$. The transformed variable R_j^i is defined as

$$R_j^i = -\Big(J(i) + 1 - j\Big) \cdot \log\left(\frac{T_{J(i)+1}^i - T_j^i}{T_{J(i)+1}^i - T_{j-1}^i}\right), \qquad j = 1, \ldots, J(i).$$

Under the null hypothesis that the bed-request rate is constant within each time interval, the $\{R_j^i\}$ are independent standard (with rate 1) exponential random variables (see the derivation in [5]). Third, we aggregate the transformed values $\{R_j^i\}$ from intervals in a certain set of days and perform the Kolmogorov-Smirnov (K-S) test on the assumption of standard exponential distribution.

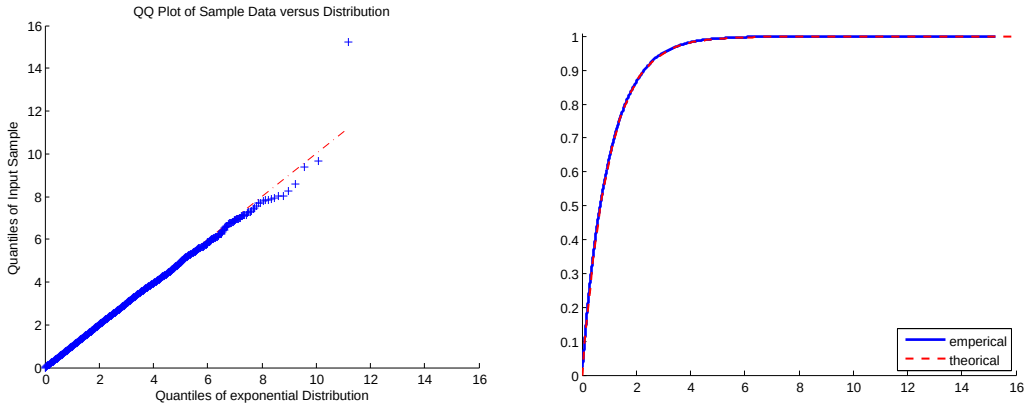
The second column in Table 8 shows the K-S test results on testing the bed-request process for each month, i.e., we aggregate $\{R_j^i\}$ from all intervals belonging to each month of Periods 1 and 2 (there are about $7 \times 30 = 210$ time intervals in a month), and perform 30 sets of K-S test for the 30 months. We can see that at significant level of 5%, 24 null hypotheses (out of 30) are not rejected.

We also perform K-S tests for longer time windows, e.g., aggregating all intervals from the 18 months in Period 1. Due to the large sample sizes (more than 35000 samples in Period 1), the p-value of K-S test at significance level 5% is very close to zero, so it is difficult to pass the test. However, the Q-Q plot and CDF plot in Figure 14 still show that the distribution of the transformed values $\{R_j^i\}$ from all intervals in Period 1 is visually close to the standard exponential distribution.

The above test results suggest that it is reasonable to assume the bed-request process from ED-GW patients to be non-homogeneous Poisson with piecewise-constant arrival rates. But note that the null hypothesis in the test does not contain any assumption on the bed-request rates of different intervals being equal or having a certain relationship. In particular, the test results do not suggest that the bed-request rate function is periodic. On the contrary, we find that the bed-request process is *not* a periodic Poisson process if using one *day* or one *week* as a period. Figures 15a and 15b clearly show that the bed-request rates depend on the day of week, so the bed-request process cannot

month	bed-request (ED-GW)	arrival (ED-GW)	EL	SDA	ICU
2008-01	0.0134	0.6800	0.0071		
2008-02	0.0638	0.8748	0.0849		
2008-03	0.0479	0.8356	0.1802		
2008-04	0.1842	0.0061	0.0178		
2008-05	0.0062	0.3073	0.0228		
2008-06	0.2150	0.1225	0.0002	0.0053	0.0000
2008-07	0.1028	0.6011	0.1148	0.0000	0.0001
2008-08	0.1949	0.7217	0.0388	0.0000	0.0000
2008-09	0.1064	0.1153	0.0253	0.0055	0.0000
2008-10	0.1253	0.2449	0.0256	0.0279	0.0003
2008-11	0.2442	0.0971	0.0026	0.0001	0.0005
2008-12	0.3092	0.7980	0.0910	0.0098	0.0000
2009-01	0.1218	0.3710	0.4210		
2009-02	0.0694	0.4445	0.0061		
2009-03	0.1860	0.1021	0.0925		
2009-04	0.1112	0.2732	0.0091		
2009-05	0.0565	0.4431	0.0729		
2009-06	0.0180	0.3783	0.1876		
2010-01	0.3259	0.5978	0.0018		
2010-02	0.9596	0.5694	0.5737		
2010-03	0.0851	0.1882	0.0007		
2010-04	0.6379	0.8514	0.0004		
2010-05	0.2684	0.0170	0.0338		
2010-06	0.0030	0.2371	0.1048	0.2959	0.0028
2010-07	0.0065	0.3571	0.4903	0.0000	0.0000
2010-08	0.0546	0.0023	0.0064	0.0103	0.0000
2010-09	0.4329	0.5033	0.4402	0.0004	0.0000
2010-10	0.7950	0.2549	0.0472	0.0485	0.0000
2010-11	0.0563	0.6290	0.0005	0.0064	0.0000
2010-12	0.1996	0.3095	0.3198	0.0127	0.0000

Table 8: Results for Kolmogorov-Smirnov test.



(a) QQ plot against the exponential assumption

(b) CDF plot against the exponential assumption

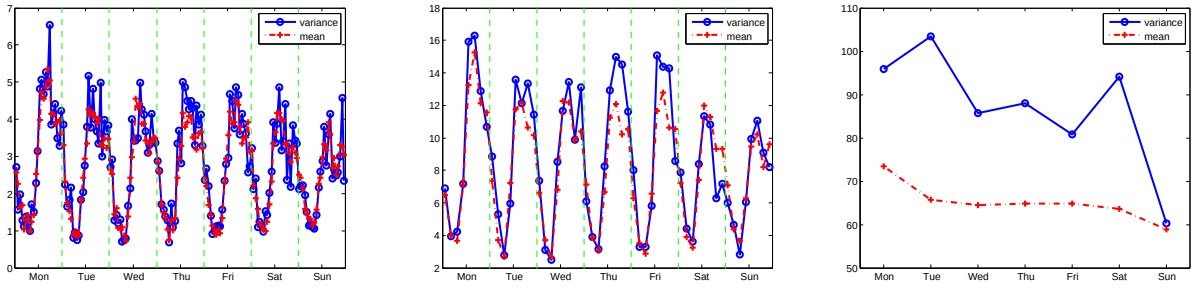
Figure 14: QQ plot and CDF plot of $\{R_j^i\}$ from all intervals in Period 1.

be periodic Poisson with one day as a period. We then examine whether the bed-request process is periodic Poisson with one week as a period. If this assumption were valid, then for each day of the week, the daily bed-request on that day in all weeks would have formed an iid sequence following a Poisson random variable. As a consequence, the mean and variance of the daily bed-request on that day of the week would be equal or close. However, Figure 15c shows that the sample variances are significantly larger than the sample means for each day of the week except for Sunday, which indicates that the bed-request process is not a periodic Poisson process with one week as a period. We conjecture that the high variability comes from the seasonality of bed-requests (e.g., February has a lower bed-requests rate than other months; see the red curve in Figure 5a)) and the overall increasing trend in the bed demand (see discussions in Section 3.3).

Figure 15c demonstrates that, under the 1-day resolution, the bed-request process shows over-dispersion, a term that was coined in Maman [16] and means that the arrival process has "significantly larger values of the sampled CV's compared to the CV's one would expect for data generated by a Poisson distribution." Unlike the 1-day resolution case, we observe from Figures 15a and 15b that, under the 1-hour and 3-hour resolutions, the sample means and sample variances are close for most intervals. This observation is consistent with the findings in Section 3.3 of [16] and suggests that variability of bed-request rates at these two resolutions is close to (or somewhat larger than) the variability of iid Poisson random variables. Note that we have differentiated among seven days in a week in Figures 15a and 15b to account for the day-of-week variations; Maman [16] did the same when testing the arrival process to ED (see Section 3.3 in her paper). If we do not differentiate, the over-dispersion phenomenon would be more prominent. Maman [16] also gave a possible explanation for the phenomenon that the difference between the empirical and Poisson CV's increases when one decreases the time resolution (see Remark 3.3 there).

For the arrival process to ED from ED-GW patients (see the blue curve in Figure 12b), we perform similar tests. The third column of Table 8 shows the K-S test results on the non-homogeneous Poisson assumption for the arrival process to ED for each month in Periods 1 and 2. Not surprisingly, more null hypotheses (27 out of 30) are not rejected. We observe similar phenomena on over-dispersion under the three time-resolutions tested in Figure 15.

Finally, we perform the tests on the non-homogeneous Poisson assumption for bed-request processes from SDA and ICU-GW patients and admission process from EL patients (i.e., using EL patient's admission time stamp). The fourth to sixth columns of Table 8 show the K-S test results



(a) 1-hour resolution (b) 3-hour resolution (c) 24-hour resolution

Figure 15: Comparison between sample means and sample variances of bed-requests in 1-hour, 3-hour and 24-hour resolutions using Period 1 data.

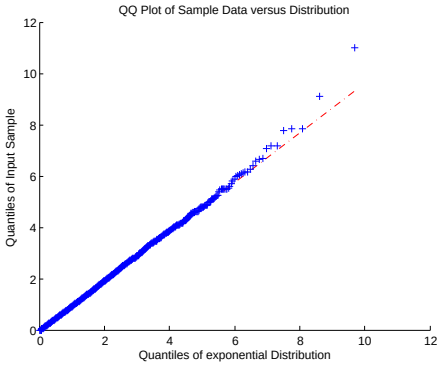
for each month in Periods 1 and 2. Note that we only have 14-month data for the bed-request times of SDA and ICU-GW patients (see explanation in Section 2.5). Thus, the last two columns of Table 8 only display the K-S test results for the corresponding months. From the table, we see at significant level of 5%, 17 null hypotheses out of 30 are rejected for the EL admission process, and nearly all the null hypotheses are rejected for SDA and ICU-GW bed-request processes (13 and 14, out of 14, are rejected for SDA and ICU-GW, respectively). Similar to Figure 14, Figure 16 shows the Q-Q plots and CDF plots for the EL admission process and for the SDA and ICU-GW bed-request processes. In the figure, the transformed values $\{R_j^i\}$ from all intervals in Period 1 are aggregated. We observe that the distribution of the transformed values for EL admission process is still visually close to the standard exponential distribution, but not for the other two tested processes.

7 Length of Stay

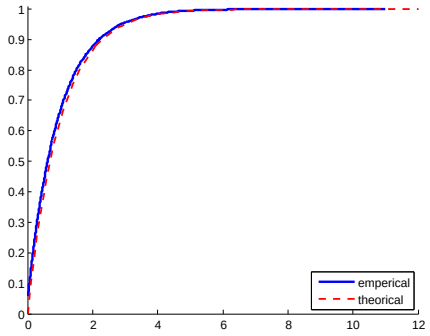
The length of stay (LOS) of an inpatient is defined as the number of nights the inpatient stays in the hospital. For a patient who has transferred at least once during the entire hospital stay, we also define her LOS in a certain ward as the number of nights she stays in that particular ward (see more details in Section 10 and Section 4.3 of the main paper [25]). Unless otherwise specified, we use the LOS of the entire hospital stay when reporting statistics in this online supplement. Like many hospitals, NUH uses average LOS as a key performance indicator in its inpatient flow management. Note that LOS takes on only integer values. Our definition for LOS is consistent with the definition adopted by most hospitals and what is used in the medical literature. The only difference is for same-day discharge patients. According to our definition, their LOS is zero. However, for billing purposes, most hospitals treat these patients as having stayed one day (i.e., adjust their LOS to 1). See, for example, the National Hospital Discharge Survey (NHDS) [7, 11]. Since our paper focuses on hospital operations, we ignore the adjustment for same-day discharge patients used by NHDS.

LOS is not the same as *service time*, which refers to the duration between patient admission and discharge. LOS, we believe, depends mainly on the patient's medical conditions, and is less sensitive to many operational policies such as the discharge policy. In some studies (e.g., [3]), LOS and service time are used synonymously. In the main paper [25], the authors demonstrate that it is important to work with LOS distributions, and not directly with service time distributions to build high-fidelity operational models for inpatient flow management. Thus, we need to differentiate these two notions.

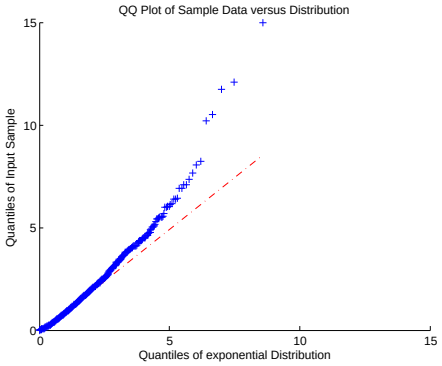
In Section 7.1, we present the LOS distributions for all general patients in Periods 1 and 2. In



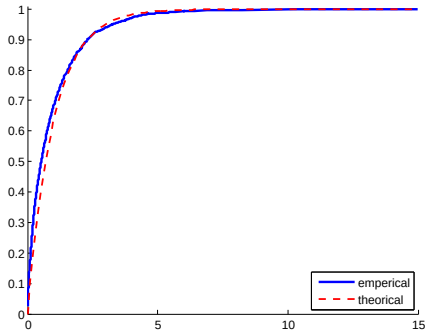
(a) QQ plot (EL admission)



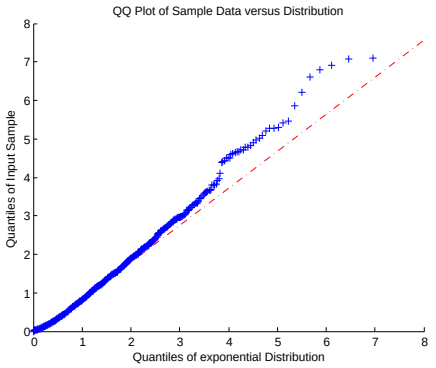
(b) CDF plot (EL admission)



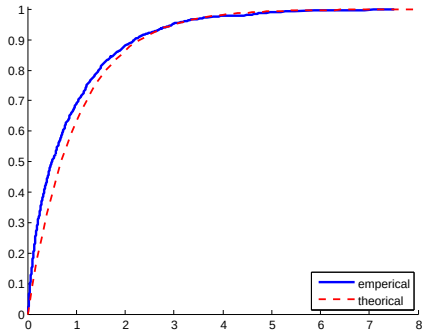
(c) QQ plot (ICU-GW bed-request)



(d) CDF plot (ICU-GW bed-request)



(e) QQ plot (SDA bed-request)



(f) CDF plot (SDA bed-request)

Figure 16: QQ plots and CDF plots (against the standard exponential assumption) of $\{R_j^i\}$ from all intervals in Period 1 for the admission process of EL patients and the bed-request processes of ICU-GW and SDA patients.

LOS	Period 1	Period 2	LOS	Period 1	Period 2
0	2.85%	3.30%	16	0.52%	0.43%
1	19.99%	20.40%	17	0.44%	0.32%
2	21.62%	22.05%	18	0.32%	0.30%
3	14.85%	14.95%	19	0.32%	0.25%
4	9.99%	10.20%	20	0.29%	0.26%
5	6.86%	6.88%	21	0.26%	0.21%
6	5.05%	4.98%	22	0.23%	0.20%
7	3.69%	3.43%	23	0.15%	0.17%
8	2.75%	2.69%	24	0.18%	0.21%
9	2.08%	1.96%	25	0.12%	0.11%
10	1.70%	1.51%	26	0.14%	0.13%
11	1.29%	1.05%	27	0.07%	0.08%
12	1.10%	0.93%	28	0.10%	0.08%
13	0.85%	0.82%	29	0.07%	0.08%
14	0.70%	0.67%	30	0.09%	0.05%
15	0.55%	0.56%	>30	0.78%	0.73%

Table 9: LOS distribution (cut-off at 30 days).

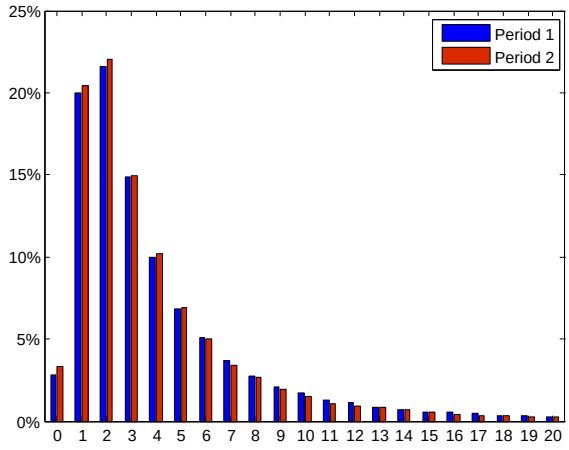
Section 7.3, we demonstrate that the LOS distribution depends on patient admission sources and patient specialities. In Section 7.2, we discuss the rationale for dividing ED-GW patients based on admission time and compare the LOS distributions for AM-patients and PM-patients. In Section 7.4, we investigate the effect of wrong ward assignment on patient LOS by comparing the average LOS between overflow patients and right-siting patients.

Note that in this section, all the LOS distributions are for patients who did *not* transfer between general wards and ICU-type wards during their stays. Patients who have transferred during the stay show different LOS distributions. See Section 10.3 for more empirical results. The proposed stochastic model in [25] uses patient classes to differentiate LOS distributions between transfer and non-transfer patients.

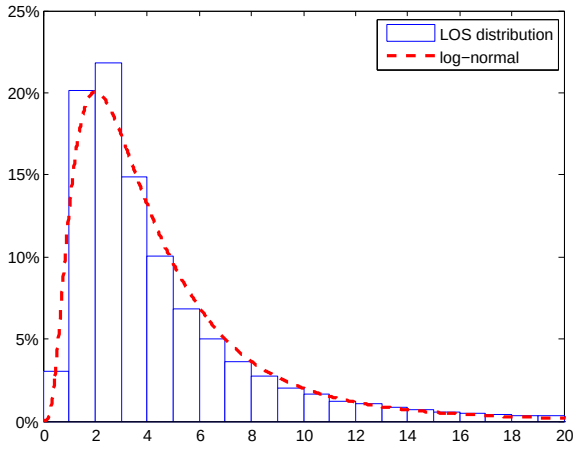
7.1 LOS Distribution

Table 9 lists the empirical distributions of LOS in Periods 1 and 2, with the cut-off value at 30 days. The means (without truncation) for Periods 1 and 2 are 4.55 and 4.37 days, respectively. The coefficients of variations (CVs), which is defined as the standard deviation divided by the mean, are 1.28 and 1.29, respectively. More than 95% of the patients have LOS between 0 and 15 days in both periods. Figure 17a plots the LOS distributions in two periods with the same cut-off as in Table 9. The two distributions are both right-skewed. About 0.78% and 0.73% of the patients stay in NUH for more than 30 days in Periods 1 and 2, respectively, although the average LOS is only about 4.5 days for both periods. The maximum LOS is 206 days for Period 1 and 197 days for Period 2. Table 10 lists the tail frequencies of LOS after 30 days for the two periods. The bin size is 5 days and the cut-off value is 90 days.

Figure 17a and Tables 9 and 10 show little difference in the LOS distributions between Periods 1 and 2. We now combine the data of the two periods. Figure 17b plots the empirical LOS distribution curve from the combined data, which visually resembles a log-normal distribution (with mean 4.65 and standard deviation 4).



(a) LOS distributions in two periods



(b) Fitting the LOS distribution with a log-normal distribution (mean 4.65 and std 4)

Figure 17: LOS distributions.

bin	Period 1	Period 2	bin	Period 1	Period 2
(30,35]	0.25%	0.27%	(60,65]	0.03%	0.02%
(35,40]	0.16%	0.15%	(65,70]	0.02%	0.02%
(40,45]	0.10%	0.09%	(70,75]	0.01%	0.01%
(45,50]	0.08%	0.06%	(75,80]	0.01%	0.01%
(50,55]	0.04%	0.05%	(80,85]	0.01%	0.01%
(55,60]	0.04%	0.02%	(85,90]	0.01%	0.01%
			>90	0.02%	0.02%

Table 10: LOS tail frequencies (start from 31 days, cut-off at 90 days).

7.2 AM- and PM-patients

Empirical evidence suggests that ED-GW patients' LOS depends on admission times. Figure Figure 18a plots the average LOS for ED-GW patients admitted during each hour (using combined data). We observe that patients admitted before 10am have similar average LOS, and so are patients admitted after 12 noon. There is also a spike from 10am to noon. Given these interesting features, we categorize ED-GW patients into two groups: those admitted before noon, and those admitted after noon. For convenience, from now on we refer to them as ED-AM patients and ED-PM patients, respectively.

Figure 18b provides the admission time distributions of the four admission sources. Around 69% of the ED-GW patients, 95% of the EL patients, 94% of the ICU-GW patients, and 92% of the SDA patients are admitted after noon. This suggests that for the purpose of comparing the differences of LOS between AM and PM admissions, we should focus on ED-GW patients, since patients from other sources comprise a very small portion of those admitted before noon. Moreover, in Section 7.3 we will see that EL and ICU-GW patients have longer average LOS than ED-GW patients; thus including them introduces bias in the comparison between AM and PM admissions. The sample sizes in the following analysis only include ED-GW patients.

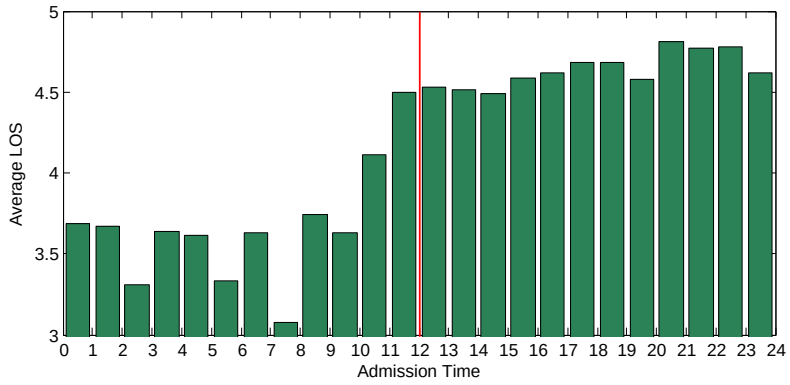
The LOS distributions for ED-AM patients and ED-PM patients are substantially different. Table 11 lists the total sample sizes and the LOS distributions, truncated to the first 21 values, for ED-AM and ED-PM patients in the two periods. Figure 19a shows the corresponding plots. The number of ED-PM patients is 2.2 times that of ED-AM patients. Around 11% to 13% of the ED-AM patients are same-day discharge patients (i.e., those with LOS=0), whereas nearly 0% of the ED-PM patients are discharged same day in the two periods.

Close examination reveals a difference of about one day between the average LOS for ED-AM and ED-PM patients. Using combined data, the average LOS is 3.60 days for all ED-AM patients and 4.66 days for all ED-PM patients. In fact, the two LOS distributions in Figure 19a are similar in shape when we do a shift. Figure 19b shows the comparison between the LOS distribution for ED-PM patients and the *shifted* LOS distribution for ED-AM patients. Here the shifted distribution means that we shift the LOS distribution to the right-hand side of x-axis by 1. For example, value 1 in this plot corresponds to value 0 in the original LOS distribution for ED-AM patients. We omit ED-PM patients with LOS=0 in Figure 19b due to the negligible proportion, so the plots start from value 1. After the shift, the two distribution curves are indeed close. The one-day difference in average LOS between ED-AM and ED-PM patients persists when we look into each specialty. See Section 7.3 for more details.

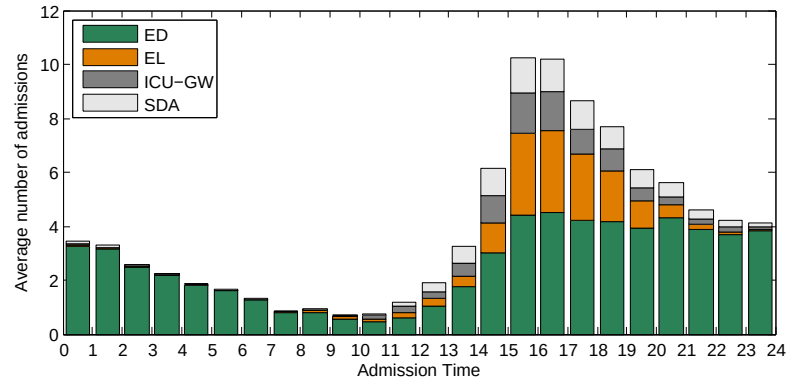
We speculate a potential reason for the one-day difference between ED-AM and ED-PM patients is practicality, i.e. most tests, consulting, and treatment occur between 7am and 5pm (the regular working hours). ED-AM patients can therefore be subjected to these tests and treatment since most of them are admitted in early morning (before 6am), whereas ED-PM patients must wait until the following day since most admissions are after 4pm. We use the following example with two scenarios for further illustration. In this example, We make three assumptions: an AM patient is admitted at 2am; a PM patient is admitted at 4pm; and both patients are discharged at 3pm. These assumptions actually represent a typical situation, since most ED-AM patients are admitted between midnight and 4 am, most ED-PM patients are admitted between 3pm and 8pm, and the discharge peak is between 2pm and 3pm.

Scenario 1

An AM-patient admitted at 2am on May 1, 2008 has a medical condition that requires 1 day for surgery and 2 days for pre/post-surgery testing and treatment. She can utilize the day of admission



(a) Average LOS for ED-GW patients admitted in each hour

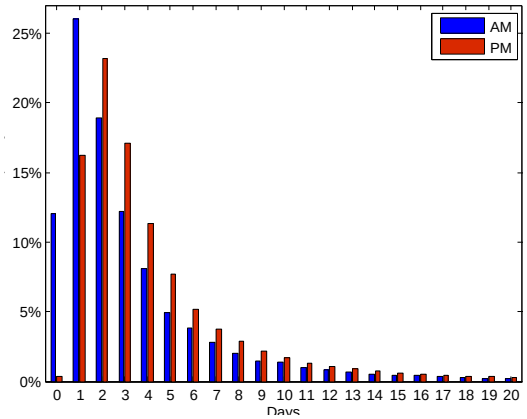


(b) Average number of admissions for ED-GW, EL, ICU-GW, and SDA patients in each hour

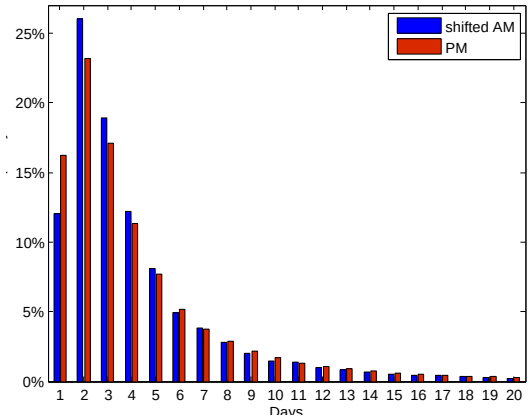
Figure 18: Average LOS with respect to the admission time (combined data).

	ED-AM		ED-PM	
	Period 1	Period 2	Period 1	Period 2
	sample size			
	10156	7189	22897	16046
LOS	distribution (%)			
0	11.29	13.12	0.32	0.42
1	25.67	26.44	15.45	17.33
2	18.87	18.97	23.03	23.44
3	12.40	11.95	17.02	17.23
4	8.04	8.07	11.38	11.25
5	5.18	4.49	7.77	7.59
6	3.92	3.78	5.28	4.97
7	2.88	2.66	3.75	3.66
8	1.93	2.18	3.00	2.78
9	1.57	1.36	2.25	2.14
10	1.50	1.13	1.81	1.56
11	1.03	0.95	1.42	1.06
12	0.94	0.64	1.15	0.98
13	0.70	0.61	0.92	0.83
14	0.49	0.49	0.76	0.72
15	0.46	0.40	0.57	0.56
16	0.48	0.29	0.48	0.52
17	0.36	0.26	0.54	0.33
18	0.27	0.22	0.36	0.29
19	0.22	0.22	0.36	0.26
20	0.18	0.21	0.31	0.26
>20	1.63	1.54	2.08	1.82
average	3.70	3.46	4.78	4.48

Table 11: LOS distributions for ED-AM and ED-PM patients; sample sizes only include ED-GW patients.



(a) Original LOS distribution for ED-AM and ED-PM patients



(b) LOS distribution for ED-AM patients is shifted to the righthand side of x-axis by 1

Figure 19: LOS distribution for ED-AM and ED-PM patients.

(May 1) to do pre-surgery tests. She receives surgery and other treatment on May 2 and May 3. She discharges at 3pm on May 4, 2008.

Scenario 2

A PM-patient admitted at 4pm on May 1, 2008 has the same medical condition as the AM-patient. But her admission time renders the day of admission wasted, and "pushes" the surgery and all pre/post testing and treatment one day later. Thus, she discharges at 3pm on May 5, 2008.

It is easy to calculate that the AM patient's entire service time is 3.54 days (85 hours) and and the LOS is 3 days, whereas the PM patient's entire service time is 3.96 days (95 hours) and the LOS is 4 days. The difference between the service time in the two scenarios is 0.42 day (10 hours), and the difference between LOS is 1 day. All these numbers match the statistics we show for ED-AM and ED-PM patients (see Section 8.3 for statistics on service time); thus, our explanation is reasonable.

7.3 LOS distributions according to patient admission source and specialty

Table 12 reports the average and standard deviation of the LOS for each specialty and for each admission source in Periods 1 and 2. From the table, we can clearly see that the average LOS is both admission-source and specialty dependent. Moreover, consistent Section 7.2, the one-day difference in average LOS between ED-AM and ED-PM patients exists across all specialties.

Combining the data points for the two periods, we plot the LOS distributions for the 9 major specialties and the four admission sources in Figure 20. From Table 12 and the figures, we observe the following:

1. Comparing among specialties, Oncology, Orthopedic and Renal patients record a longer average LOS. Surgery and Cardiology patients demonstrate a shorter average LOS. The LOS distributions of each specialty exhibit a similar shape, which resembles a log-normal distribution. Oncology and Renal patients tend to have a longer tail. Both have a high proportion of patients staying longer than 14 days (9.93% for Oncology, and 7.59% for Renal, compared with

Cluster	Period	ED-GW(AM)	ED-GW(PM)	EL	ICU-GW	SDA
Surg	1	2.36 (2.93)	3.27 (3.43)	4.55 (6.55)	9.58 (12.60)	2.59 (4.72)
	2	2.37 (3.04)	3.25 (3.40)	4.71 (6.11)	10.12 (13.32)	3.63 (8.09)
Card	1	2.95 (3.75)	3.83 (3.93)	4.15 (5.08)	5.22 (6.78)	2.55 (3.38)
	2	3.02 (3.93)	4.01 (4.68)	4.15 (5.64)	5.15 (7.47)	2.75 (4.26)
Gen Med	1	3.94 (4.76)	5.25 (5.87)	5.32 (5.79)	10.43 (18.43)	3.17 (2.62)
	2	4.09 (5.41)	5.24 (5.35)	5.47 (6.20)	8.82 (13.69)	3.15 (2.26)
Ortho	1	5.45 (8.22)	6.04 (7.04)	6.27 (6.19)	10.82 (13.32)	3.41 (4.32)
	2	3.27 (4.52)	4.65 (5.64)	6.15 (7.04)	13.49 (13.82)	4.62 (6.49)
Gastro	1	3.32 (3.91)	4.48 (4.47)	3.70 (4.39)	8.33 (12.25)	3.24 (3.99)
	2	3.51 (6.14)	4.18 (5.10)	3.55 (3.32)	6.97 (8.76)	3.27 (5.24)
Onco	1	5.93 (7.58)	7.03 (7.14)	6.45 (7.95)	8.62 (9.02)	4.10 (4.18)
	2	5.56 (6.15)	6.62 (6.69)	6.32 (8.22)	7.65 (9.06)	4.38 (5.40)
Neuro	1	3.23 (5.22)	4.07 (4.69)	4.06 (4.69)	7.56 (7.67)	2.59 (2.40)
	2	2.98 (6.69)	3.51 (4.52)	4.50 (4.77)	9.16 (11.85)	2.45 (1.85)
Renal	1	5.75 (6.55)	6.51 (6.90)	5.70 (6.20)	10.22 (12.91)	2.08 (1.16)
	2	4.63 (6.56)	5.40 (6.01)	5.06 (5.80)	8.65 (12.20)	3.30 (3.27)
Respi	1	3.21 (5.10)	4.29 (4.26)	4.45 (6.27)	7.86 (10.71)	2.33 (3.33)
	2	2.89 (3.65)	4.28 (4.27)	3.68 (3.81)	7.36 (9.70)	3.43 (2.07)
All	1	3.70 (5.25)	4.78 (5.45)	5.17 (6.47)	7.59 (10.82)	2.84 (4.29)
	2	3.46 (5.10)	4.48 (5.11)	5.11 (6.57)	7.62 (10.77)	3.66 (6.63)

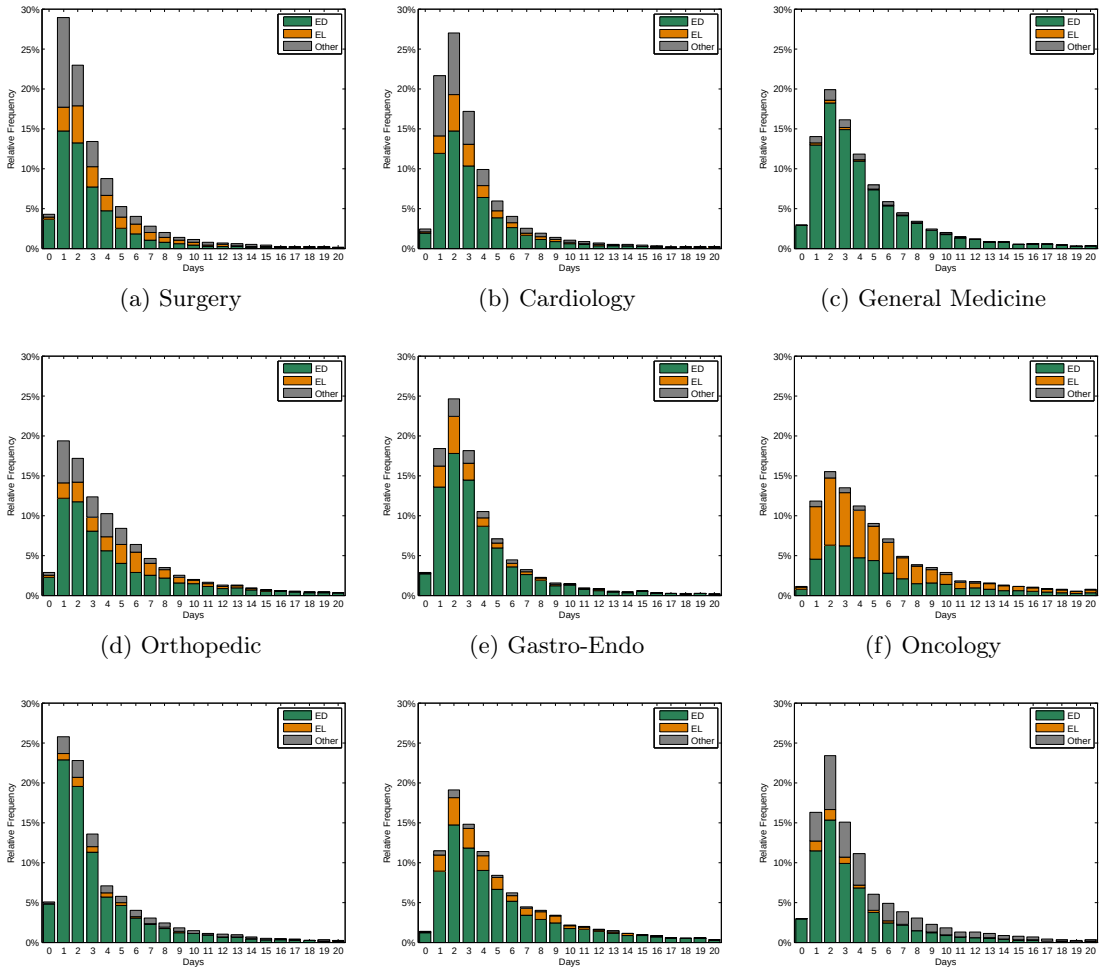
Table 12: Average LOS in days for patients in each specialty from four admission sources; number in parentheses is the corresponding standard deviation.

4.95% for all patients). The Coefficients of Variation (CV) for most combinations of specialty and admission source are between 1 and 2 in both periods. ICU-GW patients from specialties belonging to the Medicine cluster show a large CV (e.g., General Medicine, Respiratory), due to their small sample sizes.

2. Comparing across all admission sources, in general, SDA patients have a shorter average LOS (about 2-3 days); ICU-GW patients, however, have a much longer average LOS than patients from other sources for most specialties. Comparing EL and ED-GW patients shows that EL patients tend to have a longer average LOS than ED-GW for most specialties.

3. Comparing the average LOS for the two periods shows large differences in certain specialties, although we observe little difference for all patients between the two periods (see Figure 17a). Renal patients show a significant decrease in average LOS for all admission sources except SDA patients (a reduction of about 1 day) in Period 2. Orthopedic also shows a significant reduction in the average LOS for ED-GW patients. Other specialties show similar average LOS for the two periods. To further compare the differences in the two periods, Figure 21 compares the LOS distribution curves for Renal and Orthopedic in each period (aggregate among all admission sources). Figure 22 shows the tail distributions (day 11 to day 30) for these two specialties in each period. We observe a reduction in the tail distribution in the second period for both Renal and Orthopedic. This indicates that fewer patients stay at NUH for more than 20 days.

The heterogeneity of the average LOS among specialties is expected, since the underlying medical conditions for patients of the different specialties are markedly different. Moreover, the patient



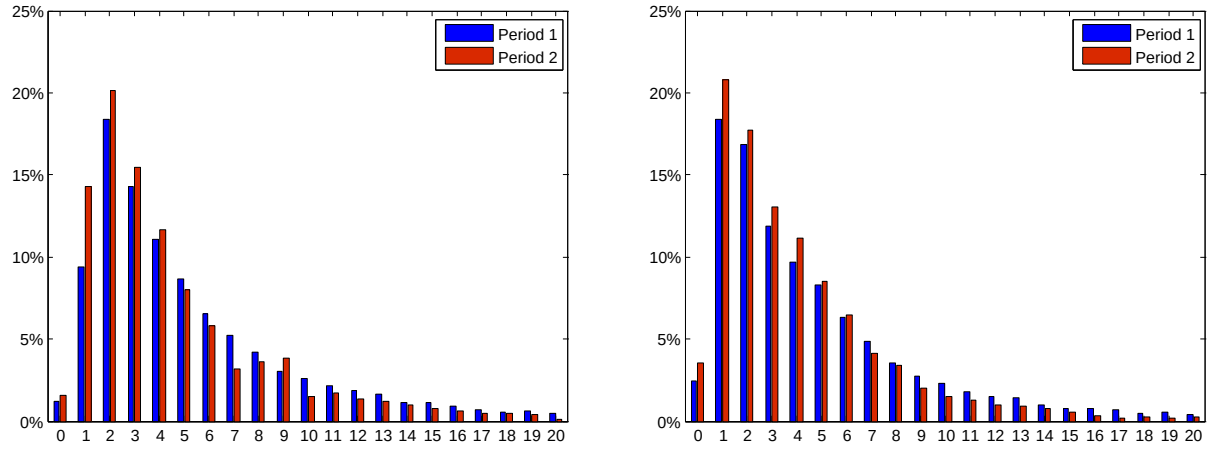
(a) Surgery (b) Cardiology (c) General Medicine

(d) Orthopedic (e) Gastro-Endo (f) Oncology

(g) Neurology (h) Renal (i) Respiratory

Figure 20: LOS distributions of the major specialties (combined data). ED-AM and ED-PM patients are aggregated under the group "ED", while ICU-GW and SDA patients are aggregated under the group "Other".

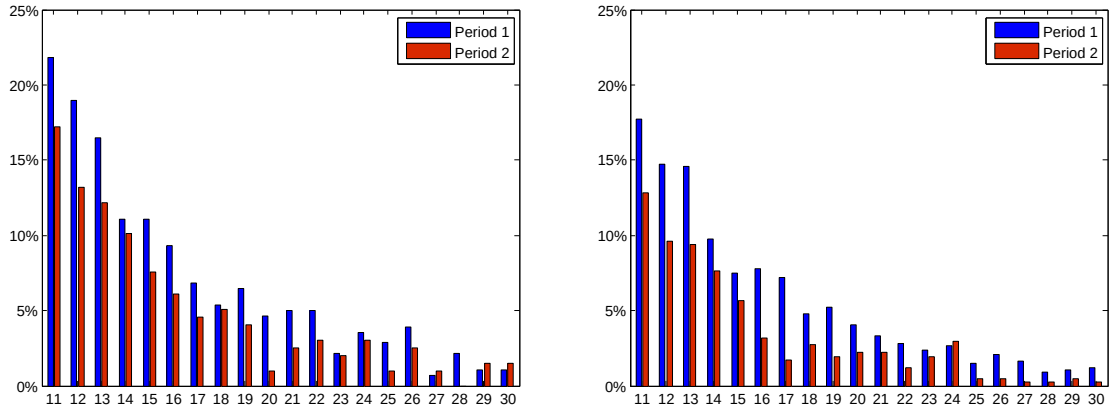
admission source also influences the average LOS. Among the specialties, we note that the average LOS of EL patients is longer than that of ED-GW patients for Surgery, Cardiology, and Orthopedic. This is somewhat counter-intuitive, since ED-GW patients generally have more urgent and complicated conditions than EL patients and need longer treatment time. One possible explanation is that most EL patients (from these specialties) undergo surgical procedures during their stay, but their priority in surgery scheduling is lower than that of ED-GW patients. EL patients usually are admitted at least one day earlier before the day of surgery, while ED-GW patients may have their surgeries done on the same day of admission due to the urgency. We note that hospitals in other countries report similar dependency of average LOS on admission sources (ED-admitted or elective), e.g., UK [20], although some also report shorter averages for elective patients, e.g., Canada (see Page 14 of [1]) and US [6, 13]. The difference could probably be the result of financial incentives and related factors in place.



(a) Renal

(b) Orthopedic

Figure 21: LOS distributions of Renal and Orthopedic patients in Periods 1 and 2.



(a) Renal

(b) Orthopedic

Figure 22: Tail LOS distributions (between day 11 and day 30) for Renal and Orthopedic patients in Periods 1 and 2.

7.4 LOS between right-siting and overflow patients

As introduced in Section 5, NUH sometimes overflows patients to non-primary wards. We call a patient who is assigned to a non-primary ward an *overflow* patient, otherwise a *right-siting* patient. In this section, we compare the LOS between right-siting and overflow patients.

From Section 7.3 we know that LOS is both specialty and admission source dependent. Thus, to eliminate the impact of these two factors and to get a "fair" comparison between right-siting and overflow patients, we need to separate for each specialty and for each admission source. Table 13 shows the comparison results, which contain the sample sizes, the average, and the standard deviation of LOS for right-siting and overflow patients. Specialties belonging to the Medicine cluster are aggregated to get a more reliable estimation (with a larger sample size).

We observe that the average LOS are close between right-siting and overflow patients for Medicine, Surgery, and Cardiology. Overflow patients from Orthopedic show a longer average LOS than that of right-siting patients for each admission source with the exception of SDA. In contrast, Oncology overflow patients show a shorter average LOS than that of right-siting patients. However, given the sample sizes of Orthopedic and Oncology overflow patients are small (as well as the high standard deviation), we cannot definitively conclude that overflow patients have a significant longer or shorter LOS than right-siting patients.

All patients in Table 13 have no transfers. In other words, the overflow patients stay in the non-primary ward until final discharge. In Section 10.2, we will discuss patients who are initially overflowed to a non-primary ward and later transferred to a primary ward. We do not include those transferred patients in the comparison here because their sample sizes are so small that we cannot separate by specialty and admission source.

8 Service times

Recall that Section 7 described the differences between LOS and service time. To construct a high-fidelity model, the main paper [25] proposes modeling service time as an *endogenous* variable, which depends on LOS, admission time, and discharge time (see Equation (3) in Section 4.3 of that paper). Since LOS constitutes the majority of a patient's service time, it is natural that service time is also admission source and specialty dependent. We do not repeat these details in this section, and aggregate patients from all admission sources and specialties in the analysis for service times. Again, we include only patients who do not transfer between GWs and ICU-type wards.

We first present some general phenomena from the service time distribution, which are summarized in Sections 8.1 to 8.3. In Section 8.5, we provide some additional empirical support for the proposed endogenous service time model in [25]. Section 4.3 of [25] compares an alternative exogenous service time model with the proposed one, and we specify the details of this alternative service time model in Section 8.4.

8.1 Service time distribution

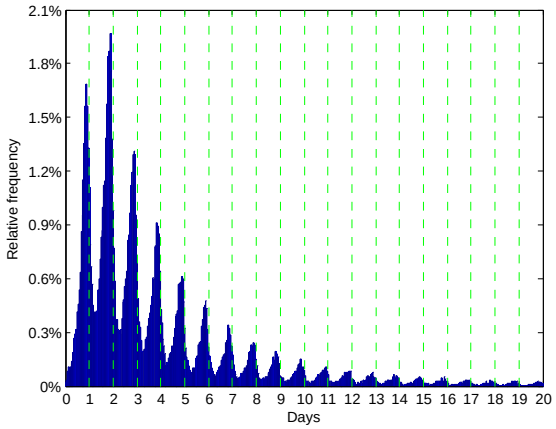
Hourly resolution

Like LOS distributions, the service time distributions for the two periods are not significantly different. Therefore, we plot them using the combined data. Figure 23a shows the histogram of the service time for all patients. The bin size is 1 hour, and each green line on the horizon axis represents a 24-hour (1 day) increment.

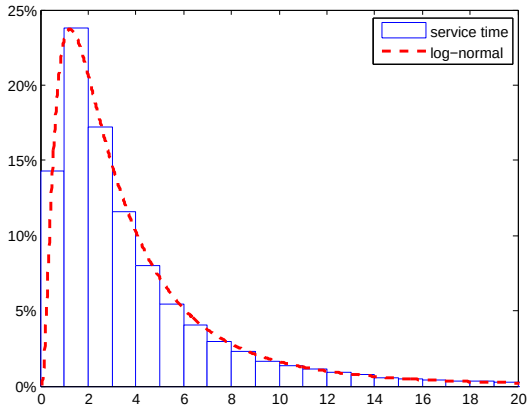
This histogram demonstrates some unique features. First, most of the data points "cluster" around the integer values (the green lines), with multiple peaks appearing at integer values which

Cluster	Source	right-siting		overflow	
		#	ALOS	#	ALOS
Med	ED-AM	2010	3.45 (4.09)	2325	3.05 (4.15)
	ED-PM	6360	4.61 (4.95)	4296	4.56 (4.87)
	EL	952	3.86 (4.52)	605	4.62 (5.50)
	ICU	756	7.08 (10.26)	779	6.64 (9.11)
	SDA	274	3.01 (2.25)	229	2.30 (1.53)
Surg	ED-AM	1364	2.23 (2.55)	537	1.85 (2.21)
	ED-PM	3040	3.04 (2.87)	590	3.04 (2.83)
	EL	1642	4.21 (6.23)	296	4.37 (5.37)
	ICU	869	7.65 (9.02)	53	8.87 (6.83)
	SDA	1894	2.26 (3.00)	281	2.12 (1.96)
Card	ED-AM	590	2.77 (3.08)	693	2.67 (3.54)
	ED-PM	1653	3.70 (3.65)	1550	3.57 (3.62)
	EL	710	3.70 (3.65)	509	4.16 (5.25)
	ICU	1332	3.95 (4.73)	237	4.27 (3.98)
	SDA	459	1.92 (1.97)	249	2.08 (2.38)
Ortho	ED-AM	971	4.62 (6.18)	155	6.55 (10.57)
	ED-PM	2363	5.53 (6.41)	488	6.46 (7.93)
	EL	1041	5.57 (4.90)	195	7.59 (7.55)
	ICU	62	8.53 (9.59)	19	11.63 (18.79)
	SDA	906	3.17 (3.03)	139	2.96 (2.42)
Onco	ED-AM	249	5.69 (7.35)	73	2.99 (2.83)
	ED-PM	645	6.81 (7.15)	171	4.09 (4.15)
	EL	1348	6.10 (7.64)	214	4.35 (7.10)
	ICU	148	7.43 (6.53)	19	6.89 (8.90)
	SDA	7	3.43 (2.64)	2	1.50 (0.71)

Table 13: Average LOS for right-siting and overflow patients using Period 1 data; numbers in parentheses are standard deviations. Only patients without any transfer activities are included. Specialties belonging to the Medicine cluster are aggregated for a larger sample size.



(a) Distribution, in hourly resolution, of the service times from both periods; each green dashed line corresponds to a 24-hour increment



(b) Distribution, in daily resolution, of the service times from both periods; a log-normal distribution fits the histogram (mean 5.02, std 6.32)

Figure 23: Distribution of the service times in two time resolutions.

represent Day 1, Day 2, ... In fact, such clustering phenomena in the service time distribution have been observed in other hospitals using the same 1-hour time resolution; see, for example, [3]. Second, we note that connecting the peak points gives a curve with a shape similar to the LOS distribution in Figure 17a. This indicates that the service time and LOS have a close relationship, although they are two different notions. See Equation (3), Section 4.3 of [25] for the relationship.

Daily resolution

Figure 23b plots the histogram of the service times using the combined data, but in daily resolution, i.e., the bin size is 1 day. Like the LOS distribution, this plot resembles a log-normal distribution, which is consistent with the observations from [3].

To populate the service time model proposed in [25], LOS distributions are estimated empirically from NUH data (see Section 7 of this document). It is tempting to use service time distributions, in daily resolution, to approximate or replace the LOS distributions. Figure 24 shows that the LOS distribution and the day-resolution service time distribution can be significantly different. Thus, it is important to estimate the LOS distributions directly from a hospital data set, rather than relying on the corresponding service time distributions to approximate the LOS distributions. Although in some papers LOS and service time are used synonymously, we advocate differentiating between the two in order to construct a high-fidelity model.

8.2 Residual distribution

To better understand the clustering phenomenon in Figure 23, we focus on the pattern of distribution around the integer values. We use $\lfloor x \rfloor$ to denote the floor of a real number x, i.e., the largest integer value r that is smaller than or equal to x. Using the time unit of 1 day, we define the residual of service time S as

$$\text{res}(S) = S - \lfloor S \rfloor. \tag{2}$$

Figure 25a shows the histograms of the residuals in Periods 1 and 2. Clearly, the distributions are both U-shaped. In fact, in both periods, more than 65% of the residuals are located between

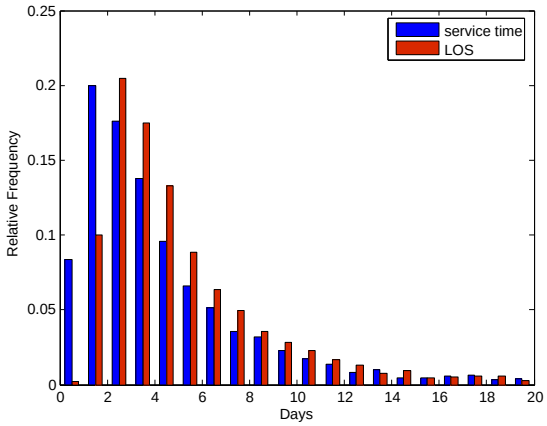


Figure 24: LOS and day-resolution service time distributions for General Medicine using the combined data.

0.58 and 1 day, and another 9% are located between 0 and .1. This U-shape results in the clustering phenomenon we observe in Figure 23. Moreover, Equation (3) below, which shows the relationship between res(S) and admission/discharge time, explains why the residual distribution has the U-shape. Let T_{adm} and T_{dis} be the admission time and discharge time of a patient, respectively (all in the unit of days). We then have

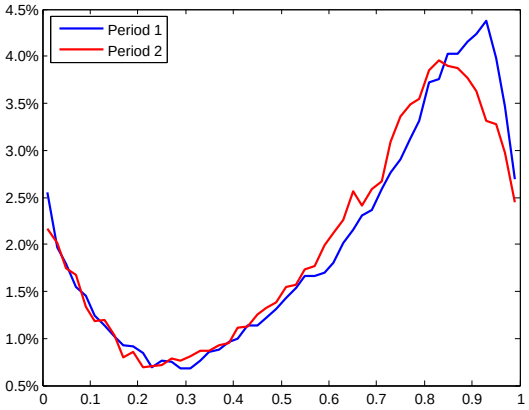
$$\begin{aligned} \text{res}(S) &= S - \lfloor S \rfloor \\ &= T_{\text{dis}} - T_{\text{adm}} - \lfloor (T_{\text{dis}} - T_{\text{adm}}) \rfloor \\ &= (T_{\text{dis}} - \lfloor T_{\text{dis}} \rfloor - (T_{\text{adm}} - \lfloor T_{\text{adm}} \rfloor)) \bmod 1, \end{aligned} \tag{3}$$

where for two real numbers x and $y \neq 0$, $x \bmod y = x - \lfloor x/y \rfloor \cdot y$.

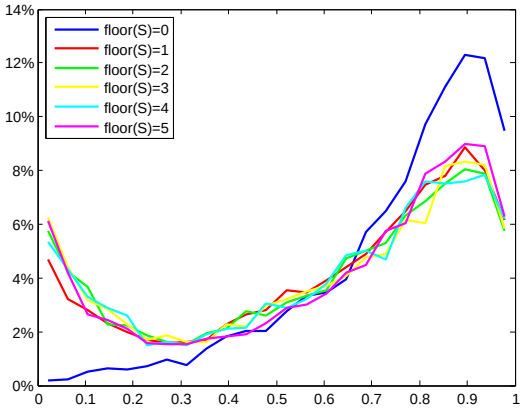
The admission and discharge time distributions jointly determine the residual distribution. We know that the majority of patients (more than 60%) are admitted between 2pm and 10pm (see Figure 18b), and discharged between noon and 4pm (see Figure 3). Thus, the "admission hour" ($T_{\text{adm}} - \lfloor T_{\text{adm}} \rfloor$) is mostly distributed between 0.58 and 0.92 day, and the "discharge hour" ($T_{\text{dis}} - \lfloor T_{\text{dis}} \rfloor$) is mostly distributed between 0.5 and 0.67 day. According to Equation (3), the residual should mostly be distributed between 0.58 and 1 day, with some distributed between 0 and 0.09. This matches Figure 25a. In summary, the "clustering" phenomenon in the service time distribution is due to the underlying admission and discharge patterns. Most admissions occur after the discharges, thus the residual is close to 0 (or 1 from periodicity).

Next, we examine whether the residual distribution depends on the value of $\lfloor S \rfloor$. Figure 25b shows the histogram of the residuals conditioning on the values of $\lfloor S \rfloor$ with Period 1 data. The bin size is 1 hour. Except for the case conditioning $\lfloor S \rfloor = 0$, the conditional residual distributions look similar and they resemble the aggregated one (the blue one) in Figure 25a. We observe the same phenomenon when we plot the conditional residual histogram using Period 2 data.

When $\lfloor S \rfloor = 0$, the conditional distribution curve is significantly different from other conditional distributions. This difference, which can also be explained using Equation (3), is mainly due to the admission and discharge distributions of same-day discharge patients (see Figure 26), which are very different from those of other patients.



(a) Empirical distribution in Periods 1 and 2



(b) Empirical distribution conditioning on floor of service times using Period 1 data

Figure 25: Empirical distribution (histogram) of the residual of service time; the bin size is 0.02 day (30 minutes).

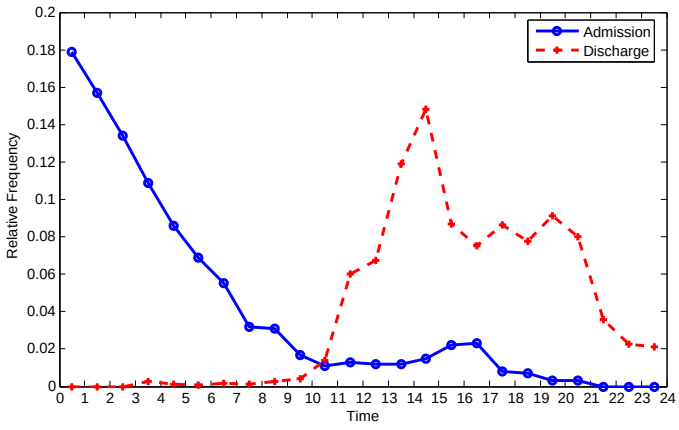
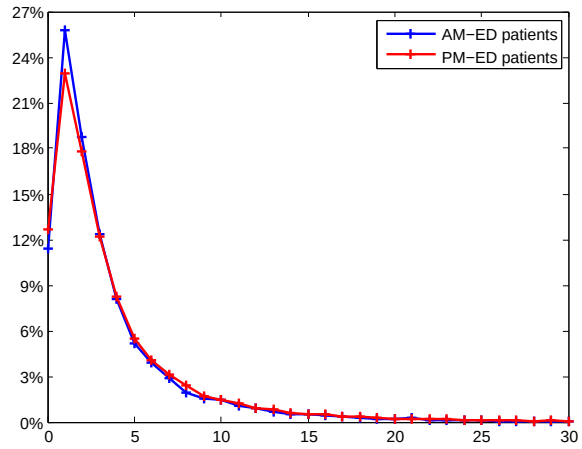
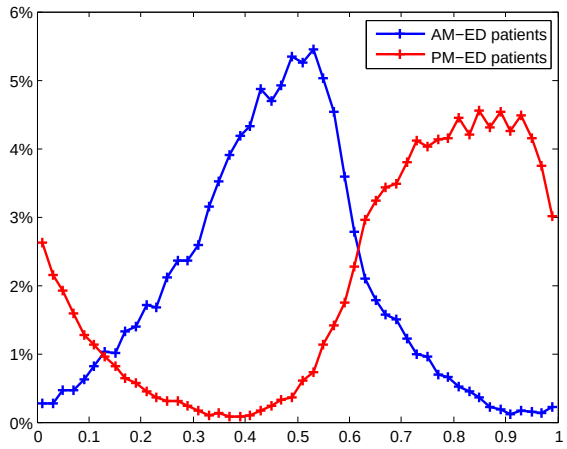


Figure 26: Admission time and discharge time distributions for same-day discharge patients using combined data.



(a) Floor of service time $\lfloor S \rfloor$; bin size is 1 day



(b) Residual of service time; bin size is 0.02 day (30 minutes)

Figure 27: Empirical distributions (histograms) of the floor and residual of service times for AM- and PM-admitted ED-GW patients using Period 1 data.

8.3 Distributions of $\lfloor S \rfloor$ and residual for AM and PM admissions

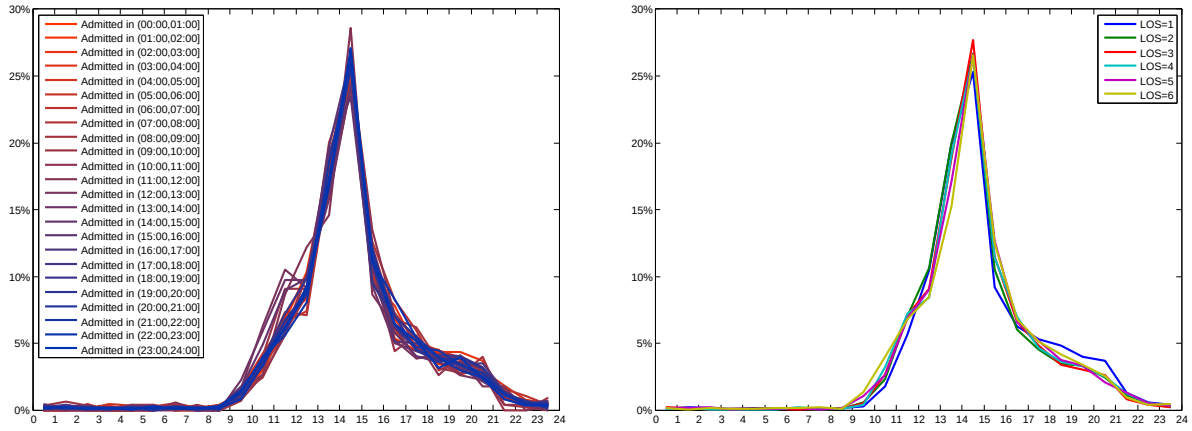
Recall that in Section 7.2, we introduce the difference between AM and PM admissions for ED-GW patients. The average LOS of AM and PM patients almost differ by 1 day; this observation holds for all ED-GW patients and for each specialty. We now investigate whether such admission-time dependency also exists in the floor of service time, $\lfloor S \rfloor$, and in the residual. Again we focus on ED-GW patients since there are few admissions before noon for patients from the other three admission sources (see discussion in Section 7.2).

Figure 27a, which compares the empirical distributions of $\lfloor S \rfloor$ for ED-AM and ED-PM patients, shows the closeness of the two distribution curves. In fact, the average of $\lfloor S \rfloor$ is 3.69 and 3.93 days for ED-AM and ED-PM patients in Period 1, respectively. The residual distributions, however, show significant differences between ED-AM and ED-PM patients (see Figure 27b). The reason can still be explained by Equation (3). The majority of ED-AM patients (around 60%) are admitted between midnight and 4am (see Figure 18b), and discharged between noon and 4pm (see Figure 3). Thus, their residuals are mostly distributed between 0.33-0.5 day, matching the blue curve in Figure 27b. For ED-PM patients, the majority are admitted between 2pm and 10pm and discharged between noon and 4pm, so the residual distribution is close to the aggregated one in Figure 25a. We get similar observations using Period 2 data.

Moreover, empirical analysis shows that the average service times are 4.15 and 3.89 days for ED-AM patients, and 4.61 and 4.30 days for ED-PM patients in Periods 1 and 2, respectively. The difference in the average service times is about 0.25 to 0.31 day (around 6-7 hours) between ED-AM and ED-PM patients, which is less than the difference in the average LOS.

8.4 Generating service times from $\lfloor S \rfloor$ and residual

Following Equation (2), one can choose to model the service time S as the sum of two random variables: an integer variable corresponding to $\lfloor S \rfloor$, and a residual variable corresponding to res(S). Moreover, Figure 25b, which shows similar distributions of the residuals regardless of the values for $\lfloor S \rfloor$, suggests an independency between the integer and residual variables. For a class of patients



(a) Discharge distribution with respect to different admission hour

(b) Discharge distribution with respect to different LOS values

Figure 28: Independence between admission and discharge hours and between LOS and discharge hours using Period 1 data.

(patient class depends on admission source, specialty, admission period, etc; see the definition in Section 3.1 of [25]), we assume that their integer and residual parts each forms an iid sequence and the two sequences are independent. Thus, the service times are also iid. This iid model is different from the non-iid service time model proposed in [25].

To populate this iid service time model, we empirically estimate the distributions for $\lfloor S \rfloor$ and $\text{res}(S)$ as shown in the previous sections. For simulation, we generate the inter and residual parts independently from the appropriate empirical distributions, and use their sum as the service time.

8.5 Additional empirical results for the service time model

The proposed service time model in the main paper is in the form of (see Equation (3) in Section 4.3 of [25]):

$$S = \text{LOS} + h_{\text{dis}} - h_{\text{adm}}, \tag{4}$$

where LOS stands for the length of stay of the patient, and h_{dis} and h_{adm} represent hour of patient admission and discharge, respectively. The model assumes that h_{dis} is independent of LOS and of h_{adm} because LOS is believed to capture the amount of time that a patient *needs* to spend in a ward due to medical reasons, whereas discharge hour h_{dis} clearly depends on the discharge patterns, which are mainly the results of scheduling and behaviors of medical staff. In this section, we provide some empirical evidence to support the assumption of the independency between h_{dis} and LOS and the independency between h_{dis} and h_{adm}. The dependency of LOS on the admission time has been discussed in Section 7.2.

Figure 28a plots the discharge distribution with respect to different admission hour, while Figure 28b plots the discharge distribution with respect to different LOS values. We note the closeness of the discharge distribution curves regardless of admission hour or LOS value. Even though we do not conduct a rigorous statistical analysis, the two figures support our assumption that the the discharge hour h_{dis} is independent of LOS and of h_{adm}.

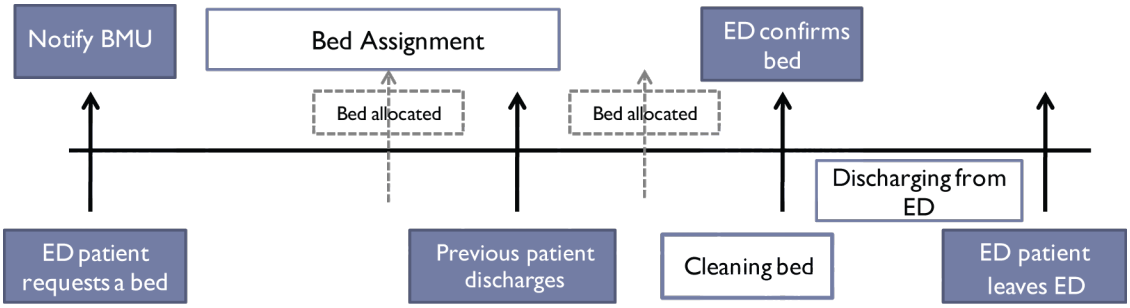


Figure 29: Process flow during the transfer from ED to GW.

9 Pre- and post-allocation delays

In [25], the authors introduce pre- and post-allocation delays, which are essential to construct a high fidelity model of the inpatient operations. The allocation delays are used to model the secondary bottlenecks besides bed availability, such as the availability of physicians, ward nurses and ED porters. The secondary bottlenecks cause additional delays during patients' admission process, i.e., even when a proper bed is vacant, a waiting patient cannot be admitted to the bed immediately. In particular, the authors show that the allocation delays have a significant impact on the waiting time of ED-GW patients. A recent study also shows that reducing the allocation delays (referred as "patient handovers" in the study) leads to a reduction in the waiting time for admission to wards [15].

Section 4.1 of [25] discusses how to model the two allocation delays and empirically estimate the parameters of their distributions. In this section, we provide a comprehensive description of the flow of a typical transfer process from the ED to a GW. It is this flow that provides the motivation of modeling pre- and post-allocation delays. We also present additional empirical analysis of the allocation delays, supplementing those already included in [25].

9.1 Transfer process from ED to general wards

When the decision is made to admit a patient, in this case an ED-GW patient, ED sends a bed-request to the bed management unit (BMU), which then initiates the search for an appropriate bed for this patient. A bed can be allocated to the patient even if it is still being occupied (but will be available soon), since BMU has "planned" discharge information. After a bed is allocated, ED confirms the allocation and then transfers the patient to the allocated bed. Figure 29 illustrates an example of the flow for transferring an ED-GW patient to a general ward. In the next two subsections, we focus on the bed allocation process and the discharge process from ED. We describe the major steps within each process, and identify the potential bottlenecks causing delays.

9.1.1 Bed allocation process

BMU controls all inpatient bed allocations at NUH during the day time, from 7am to 7pm (during the night, a nurse manager is in charge of all bed allocations). The allocation process for a bed-request from an ED-GW patient usually has four steps:

1. After BMU receives the bed request, one of the BMU staff makes a tentative bed allocation, trying to match all the criteria for the patient, such as gender, sub-specialty, class of bed, etc.

2. The staff member then checks/negotiates with the ward nurses in charge of the allocated bed in order to secure acceptance. If the ward nurses reject the request, then the staff member

makes another tentative allocation and repeats the negotiation process until one ward agrees to accept the patient.

3. Once a ward accepts the patient, BMU notifies the ED nurses about the bed allocation and waits for ED's confirmation. Occasionally the bed requirements change due to patient's medical condition, etc. Given the changed circumstances, ED cannot confirm the allocated bed and has to submit the new requirements to BMU. BMU then repeats steps 1 and 2 to effect a new allocation.

4. After ED's confirmation, the bed is officially allocated and the status of the bed displays on a screen in ED. The bed may be in different status: still occupied by the patient who is going to discharge soon, or in cleaning, or ready to be used. See more discussions below on the bed status.

The bed allocation process is similar for elective and internal transfer patients, except that when the receiving ward agrees to accept the patient, the bed assignment is confirmed via other ways (no longer through ED).

BMU has access to the status of all inpatient beds in real-time (e.g., whether a bed is currently vacant, being cleaned, or being occupied by a patient). BMU also has "planned" discharged information, which allows it to know which patients are going to be discharged. The planned discharge information also includes the ward nurses' estimate of the expected discharge time for each planned discharge patient. Therefore, BMU can allocate beds based on both the real-time status and the planned discharge. When allocation is made from planned discharge information, the bed allocation time could be even earlier than the bed available time (when the previous patient discharges). The proposed model in [25] uses two allocation modes, normal allocation and forward allocation, to capture bed allocations based on real-time and planned discharge information, respectively.

We note that the majority of time in the bed allocation process is spent on BMU staff searching for appropriate beds and negotiating with ward nurses. Insufficient number of BMU agents, especially during the peak hours when a large number of bed requests are presented (usually from 1pm to 7pm), can cause delay in the bed allocation process and thus become a bottleneck. Another bottleneck concerns ward nurse unavailability, i.e., nurses are busy with other activities (e.g., doing morning rounds with physicians) and cannot confirm BMU requests on demand.

9.1.2 Discharging from ED and transfer to wards

The ED takes the following steps to transfer ED-GW patients to wards when their allocated beds are ready:

1. ED nurses ensure that all test results are complete and no further treatment is needed in ED.

2. ED nurses check for vital symptoms to ensure patient's medical stability.

3. ED physicians give written instructions for discharge from ED.

4. ED arranges a porter (patient's escort) to transfer the patient in the company of an ED nurse.

5. Ward nurses admit the patient to the bed and "actualize" the admission.

Delays can occur in each of the above steps. For example, the patient cannot exit ED if her test results are not ready for release or she is not medically stable. Physicians or nurses may be busy attending to other ED patients, and do not have time to prepare for the ED discharge. Similarly, if ward nurses are busy, the patient cannot be admitted to her bed. Moreover, porters may not

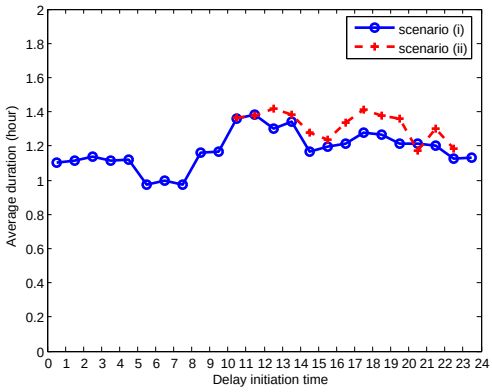


Figure 30: Estimated average of post-allocation delay with respect to the delay initiation time. Scenario (i): the allocated bed is available before the allocation time; (ii) the allocated bed is available after the allocation time. Certain time interval of the red curve is omitted because of limited data points. Post-allocation delay equals the duration between the bed allocation time and the admission time for scenario (i); and duration between the bed available time and the admission time for scenario (ii). See details in Section 4.1 of [25].

be available, especially during peak hours. We see that ED physicians, nurses, and porters may all become bottlenecks during the ED exit and transfer process.

Rationale for modeling allocation delays

The bed allocation process and the ED discharge and transfer process are the two major processes prior to a patient's admission to her bed. As from previous sections, we note that before a bed is allocated, the delay mainly comes from the BMU side; while after the bed allocation, the delay mostly comes from the ED side. Therefore, it is reasonable to use the bed-allocation time to divide the waiting time of an ED-GW patient into two parts: the first is from bed-request time to bed-allocation time, reflecting BMU's delay; and the second from bed-allocation time to admission time, reflecting ED's delay. Under the scenario when bed availability is not the constraint, the first and the second part of the waiting time corresponds to the pre- and post-allocation delay, respectively, in the proposed model [25]. Section 4.1 of [25] elaborates the details of how to empirically estimate the allocation delay from the two parts of the waiting time.

Note that the bed cleaning process is not explicitly modeled in the allocation delays. That is, the allocation-delay model (see Section 3.2 of [25]) does not differentiate the post-allocation delay distributions between the following two scenarios: (i) the allocated bed is available before the allocation time; and (ii) the allocated bed is available after the allocation time. Here, bed being "available" indicates that the previous patient occupying the bed has been discharged. In scenario (ii), the bed needs to be cleaned after the previous patient's discharge. This assumption on the post-allocation delay is supported by our empirical results. We separately estimate the average for the post-allocation delay under scenarios (i) and (ii). Figure 30 compares the hourly average between the two scenarios. We can see the blue curve, which represents scenario (i), is close to the red curve, which represents scenario (ii). The closeness of the two curves suggests that the bed cleaning time has almost no impact on the post-allocation delay.

The observation from Figure 30 can be partially explained as follows. NUH implements an auto countdown system for bed cleaning. After a patient is discharged, the bed tracking system marks the bed as "in cleaning" and automatically counts down for 30 minutes. After 30 minutes, no matter

whether the bed is indeed cleaned or not, the system changes the bed status to "vacant", indicating it is ready to serve a new patient. The ED nurses can access the bed status information in real time. They know that the ED discharge and transfer process typically takes longer than the 30-minute cleaning time. If a patient is waiting her allocated bed to receive her, the nurses usually initiate the discharge process once the bed status changes to "in cleaning" (or shortly after the change, indicated by the fact that the red curve is slightly higher than the blue curve in Figure 30). After the bed status changes to vacant, ED can then send the patient to the allocated bed. In such a way, the auto countdown system enables the nurses to do the discharge/transfer in parallel with the bed cleaning process. This ensures that the bed cleaning time does not become a major bottleneck like those discussed in Section 9.1.2.

9.2 Additional empirical results

Distribution of pre- and post-allocation delays

Figure 13 of the main paper [25] shows that the pre- and post-allocation delays depend on when they are initiated (i.e., delay initiation time). Thus, to estimate the distributions for the allocation delays, we group patients into several sub-groups according to the delay initiation hour, so that within each sub-group, the averages of pre- or post-allocation delay for each of the aggregated hours are close. For pre-allocation delay, we create 7 sub-groups: 1-3am, 3-5am, 11am-1pm, 1pm-3pm, 3pm-6pm, 6pm-9pm, and 9pm-1am (the next day). For post-allocation delay, we create two sub-groups: 10am-2pm, and 2pm-5am (the next day). These aggregations allow a larger sample size for each sub-group. We exclude patients whose pre-allocation delay initiates between 5am and 11am, and patients with post-allocation initiation times between 5am and 10am due to the small sample sizes in these time intervals. Moreover, patients selected in the distribution estimation satisfying certain criteria. Readers are referred to Section 4.1 of the main paper and the following subsection for discussions on the criteria.

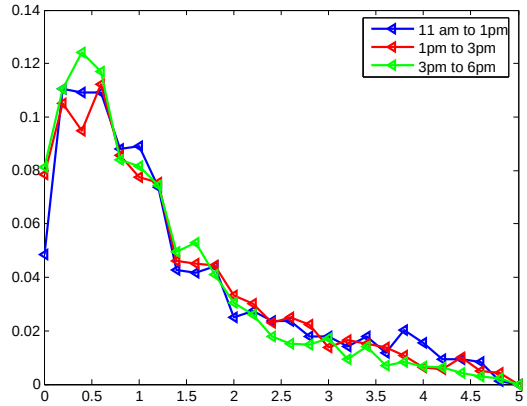
Next, we plot the histograms; Figure 31a shows them for selected pre-allocation sub-groups, and Figure 31b shows them for the two post-allocation sub-groups. We observe that all the plotted distributions resemble log-normal distributions. Plots for some other time intervals have a similar shape.

To test the reasonableness of the log-normal assumption, we perform log-transformation on the data points in each sub-group (for both allocation delays). Figures 32a and 33a show the Q-Q plot of the log-transformed data against normal distribution for the selected sub-groups. Figures 32b and 33b show the histograms of the log-transformed data and the fitted normal distributions. The figures suggest that the normal distribution curves are visually close to the empirical distribution curves. We observe similar features when analyzing the log-transformed data for other sub-groups. Although we do not conduct a rigorous statistical analysis, these figures indicate that the log-normal assumption for the pre- or post-allocation delay is reasonable and is a good starting point for building models.

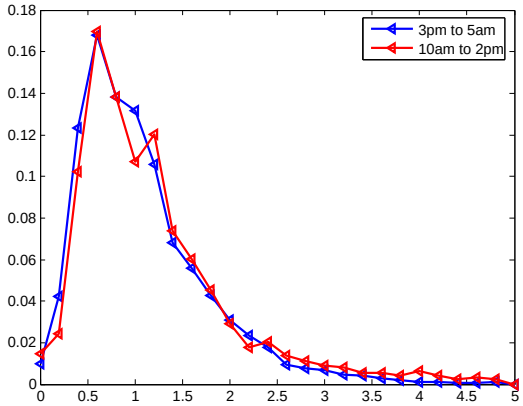
Pre-allocation delay for overflow patients

In Figure 13 of [25], the hourly average for the pre-allocation delay is estimated from patients satisfying two conditions: (i) the allocated bed is available before the bed request time; and (ii) the allocated bed comes from the primary ward for the patient. Section 4.1 of the main paper explains the reason of imposing condition (i). We now discuss the details of condition (ii).

Figure 34 compares the empirical average durations between bed-request time and bed-allocation time for right-siting and overflow patients. Condition (i) is imposed for both groups of patients, and

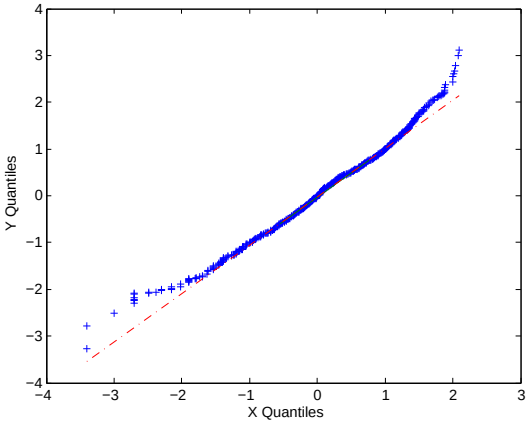


(a) Distribution of pre-allocation delay for delay-initiation time between 11 am and 1 pm, 1 pm and 3 pm, and between 3 pm and 6 pm

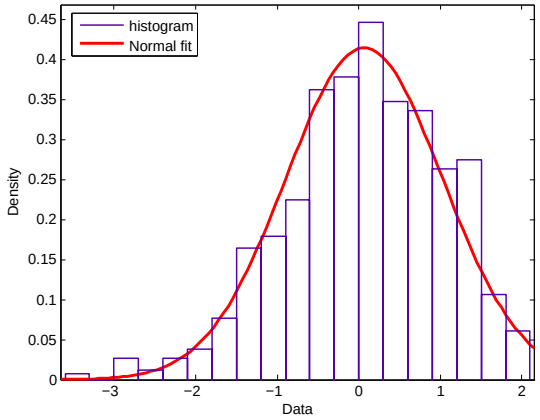


(b) Distribution of post-allocation delay for delay-initiation time between 2 pm and 5 am (next day), and between 10 am and 2 pm

Figure 31: Empirical delay distributions for bed-request in certain time intervals; bin size is 0.2 hour (12 minutes).

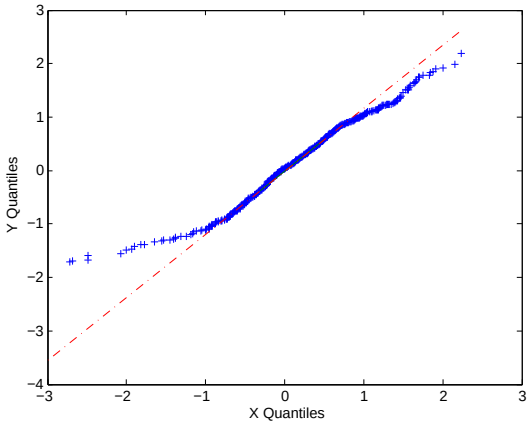


(a) Q-Q plot

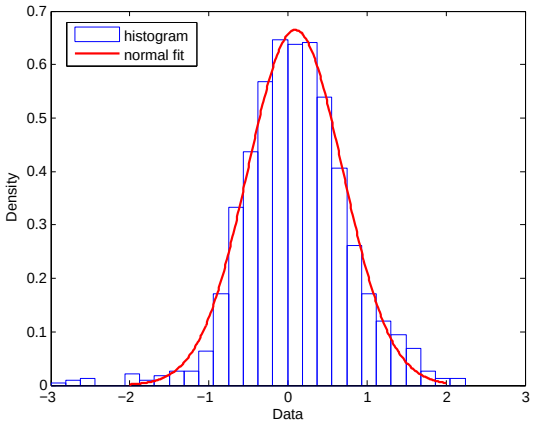


(b) Histogram and fitting with a normal distribution with mean 0.066 and std 0.96

Figure 32: Fitting the log-transformed data for pre-allocation delay with initiation time between 11 am and 1 pm.



(a) Q-Q plot



(b) Histogram and fitting with a normal distribution with mean 0.093 and std 0.60

Figure 33: Fitting the log-transformed data for post-allocation delay with initiation time between 10 am and 2 pm.

the two curves are plotted as functions of bed-request time. Clearly we can see that the average for overflow patients (red curve) is significantly longer than that for right-siting patients (blue curve). Moreover, we observe that for bed-request time from 1am to 8am, the differences in the average duration between overflow and right-siting patients are smaller than the differences in other hours.

We interpret Figure 34 with caution, because it cannot provide a definitive conclusion that overflow patients have a pre-allocation delay. In practice, BMU may wait for some time before deciding to overflow a patient if no primary bed is available upon the bed-request time. The actual search/negotiation process, which we use pre-allocation delay to capture, only starts after the overflow decision is made. Therefore, the actual pre-allocation delay for an overflow patient should equal to the duration between bed request and allocation time minus this "BMU's waiting time". However, the lack of time stamps prevents us from estimating the BMU's waiting time and thus the pre-allocation delay for overflow patients. The proposed model (see Section 5.2 of [25]) employs an overflow trigger time to mimic the BMU's waiting time, but it is only an approximation of the BMU practice and cannot be used in the allocation estimation. Thus, the proposed model in [25] does not differentiate pre-allocation delay between right-siting and overflow patients. The model estimates the pre-allocation delay distributions from right-siting patients, and use them to approximate those of the overflow patients.

Estimating the normal allocation probability $p(t)$

In Section 4.1 of the main paper [25], the authors propose using Equation (2) to empirically estimate the normal allocation probability $p(t)$ for t between 2pm and 8pm. We copy the equation here for convenience (and refer it as Equation (5)). For each hour i, we use $\hat{p}(i)$ to estimate $p(t)$ for $t \in (i, i+1]$, where

$$\hat{p}(i) = \frac{\#\text{ of patients whose allocation-completion time } > \text{ bed-available time}}{\text{total}}. \tag{5}$$

Here the total patient group consists of all ED-GW patients (in NUH data) whose bed-request time falls within that hour and whose allocated bed is not available at the bed-request time.

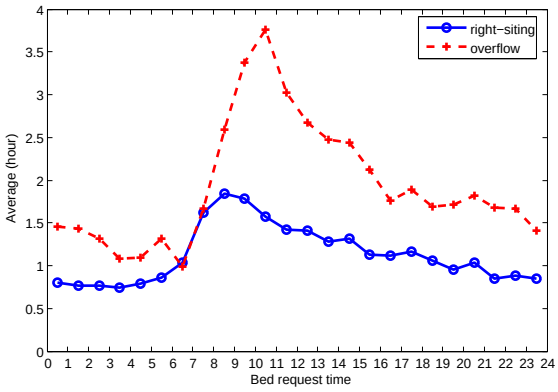


Figure 34: Average duration between bed-request time and bed-allocation time for right-siting and overflow patients. In both scenarios, the bed available time is earlier than the bed-request time.

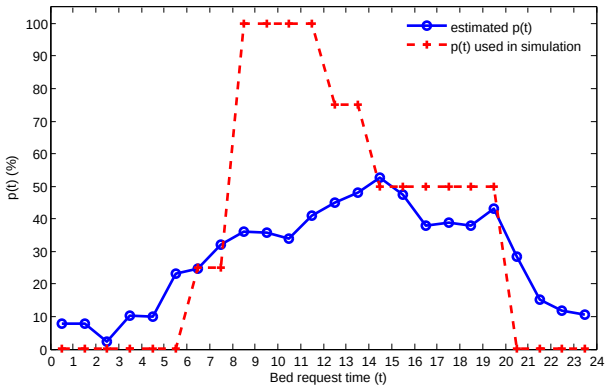
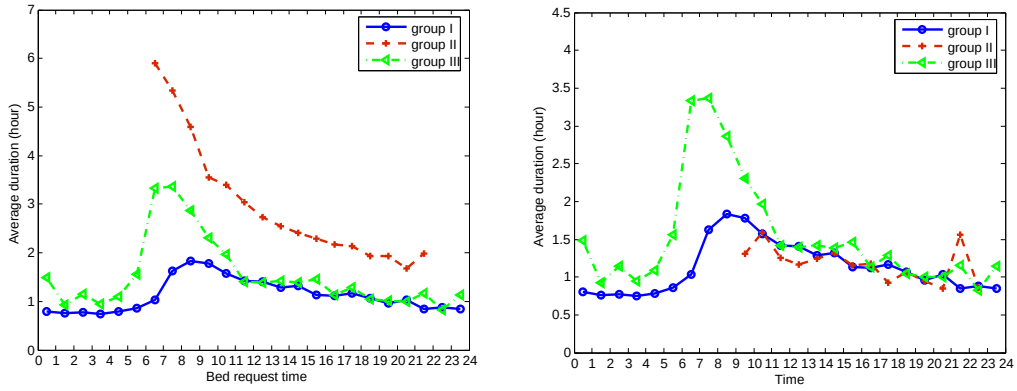


Figure 35: Estimated values of $p(t)$ from applying Equation (2) in [25], and values used in the baseline simulation. Both curves are plotted against the bed-request time t.

In this section, we provide a motivation of using this Equation to estimate $p(t)$ in certain time intervals. For that, we examine the duration between bed-request time and allocation-completion time as a function of the bed-request hour. Figure 36a plots the empirical estimate of the average of this duration among three groups of patients. The first group, corresponding to the blue curve, consists of ED-GW patients whose allocated bed is a primary bed and the bed is available at the bed-request time. The second group, corresponding to the red curve, consists of ED-GW patients whose allocated bed is not available at bed-request time and whose allocation-completion time is *later* than the bed-available time. For a patient in this group, her allocation-start time can be either before or at the bed-available time. In the latter case, the allocation is a normal allocation in our model. The third group, corresponding to the green curve, consists of ED-GW patients whose allocated bed is not available at bed-request time and whose allocation-completion time is *earlier* than the bed-available time. For a patient in this group, her allocation-start time is definitely before the bed-available time. Thus, this allocation cannot be a normal allocation. But we are not sure if it is a forward allocation because its allocation-start time may not start immediately at the bed-request time.

Back to Equation (5), one can see the patients included in the numerator is the second group of patients, while the denominator includes the second and third groups of patients. If (a) all patients in the second group have started their bed-allocation processes only at their bed-available times, and (b) all patients in the third group have started their bed-allocation processes at their bed-request times, then $\hat{p}(i)$ in Equation (5) would be a good estimator of $p(t)$. For the majority of bed-request hours, from 11am to midnight, the blue and green curves in Figure 36a are very close, suggesting that condition (b) approximately holds in this interval. To investigate condition (a), for patients represented by the red curve, we plot a modified curve in Figure 36b. In the modification, we exclude the pure waiting times due to bed unavailability, and plot the average duration between their bed-available and allocation-completion times. We can see that the modified red curve is close to the blue curve between 2pm to 8pm, suggesting that condition (a) also approximately holds in the interval. Therefore, in this time interval, it is reasonable to use $\hat{p}(i)$ to estimate $p(t)$. Outside the time interval $(12, 24)$, three curves in Figure 36b diverge, suggesting that either (a) or (b) is severely violated. Therefore, $\hat{p}(i)$ in (5) should not be used to estimate $p(t)$ outside this interval.

In Figure 35, we plot the estimated value of $p(t)$ for t in each hour from applying Equation (5). For comparison, we also plot the values used in the baseline simulation (see Equation (1) in Section 4.1



(a) Average duration for three groups of patients

(b) Revised duration for the second group

Figure 36: In (a): average duration between bed-request time and allocation-completion time for ED-GW patients as a function of bed-request time; in the red curve, we omit certain time intervals due to the lack of data points (fewer than 15 points in each hour). In (b): the red curve is a revision from the red one in (a); the two other curves are kept the same. In the revision, the duration is revised to be between bed-available time and allocation-completion time, and is plotted against the bed-available time, not the bed-request time.

of [25]). Later in Section 11.3, we perform sensitivity analysis on the choice of $p(t)$ via simulation.

10 Internal transfers after initial admission

General patients admitted from any of the four sources could go through one or more internal transfers after their initial admissions. The stochastic model developed in [25] captures the majority of these patients who have been transferred at least once. They are patients who are initially admitted to GW, moved to ICU-type wards, and then discharged (referred to as one-time transfer patients), and patients who are initially admitted to GW, moved to ICU-type wards, transferred back to GW, and then discharged (referred to as two-time transfer patients). See Sections 3.4 of [25] for details of modeling the one-time and two-time transfer patients.

In this section, we present a comprehensive empirical analysis of all patients who have gone through internal transfer(s) after the initial admissions (see Section 10.1). The empirical analysis validates our model relating to the transfer activities. In addition, we focus on another category of transfer patients not elaborated in [25] who are initially admitted to a non-primary ward or a wrong-class ward, and then transferred to a primary ward or a right-class ward (referred to as *right-siting transfers*). Section 10.2 presents the empirical analysis on these right-siting transfers. Finally, Section 10.3 shows the empirical LOS distributions for the one-time and two-time transfer patients that are captured in the stochastic model of [25].

10.1 Overall statistics on internal transfers

Using combined data, out of the total 94786 general patients admitted in both periods, 79687 (84%) have not been transferred after the initial admission. The remaining 14840 patients (16%) have gone through at least one transfer. Patients with at least a one-time transfer are termed *transfer patients*.

Specialty	ED-AM	ED-PM	EL	ICU	SDA
Surg	10.96%	9.35%	27.28%	15.18%	3.93%
Cardio	19.17%	15.94%	42.15%	11.59%	9.28%
General Med	17.80%	10.97%	15.71%	12.56%	2.94%
Ortho	14.84%	14.83%	21.03%	23.29%	5.94%
Gastro-Endo	22.19%	13.07%	8.20%	9.96%	6.64%
Onco	28.34%	23.96%	19.20%	11.21%	15.79%
Neurology	16.21%	12.23%	14.06%	12.87%	2.31%
Renal disease	30.46%	19.61%	14.78%	18.14%	6.67%
Respiratory	14.74%	14.00%	16.99%	10.09%	6.50%
Total	17.76%	13.63%	25.17%	12.62%	5.42%

Table 14: Proportion of patients who have gone through at least one internal transfer for each admission source and for each speciality using combined data.

Table 14 shows the proportion of transfer patients for each admission source and for each specialty. Clearly, the proportion of transfer patients depends on both the admission source and specialty. Generally, there are fewer SDA transfer patients. Across specialties, Cardiology and Surgery have the highest proportion of EL transfer patients, whereas Oncology and Renal have a relative higher proportion under the ED-GW source (both AM and PM admissions). We note that the proportions of transfer patients are close for AM and PM admissions (under ED-GW source) for each specialty except those belonging to the Medicine cluster. General Medicine, Renal, Neurology, and Gastro-Endo all show a higher proportion of transfer patients for AM admissions than that of PM admissions.

Overall, transfer patients admitted from ICU-GW and SDA sources make up only a small proportion of all transfer patients (1500 out of 14840, or about 10%), and account for only 1.6% of all general patients. Therefore, in the following analysis, we focus on transfer patients belonging to the ED-GW and EL sources.

Still using combined data, out of the total 77904 ED-GW and EL patients, 13340 (17%) of them have been transferred at least once after initial admission. Out of these 13340 patients, 7285 patients (54.61%) have gone through one transfer; 4428 patients (33.19%) two transfers; and 905 patients (6.78%) three transfers. The remaining 722 patients (5.41%) have been transferred more than four times, and constitute less than 0.8% of the total general patients. Therefore, we also exclude them from analysis.

Now we study the paths of these ED-GW and EL patients with one-, two-, and three-time transfers. Table 15 summarizes the information on these transfer paths. We use 1 to denote a general ward, and 0 to denote a non-general ward. Path 1-0-1 means the patient is initially admitted to a GW, transferred to a non-GW, transferred back to a GW, and then discharged.

ED-GW and EL patients with one-time transfer

Of the 7285 patients with one-time transfer, 1667 patients are transferred to a non-general ward (more than 60% to an ICU-type ward). The other 5618 are transferred to another general ward. In Section 10.2, we will study these patients transferred between two general wards in detail.

Trans times	count	path and count			
		1-0		1-1	
1	7285	1667		5618	
		1-0-0	1-0-1	1-1-0	1-1-1
2	4428	114	4036	102	176
		group I	group II	1-0-1-0	1-1-1-1
3	905	44	707	130	24

Table 15: Decomposition of ED-GW and EL transfer patients by number of transfers and pathways using combined data. In the last row, group I contains paths 1-0-0-0, 1-1-0-0, and 1-1-1-0; group II contains paths 1-0-0-1, 1-0-1-1, and 1-1-0-1.

ED-GW and EL patients with two-time transfer

Of the 4428 two-time patients, the majority (4036 patients, 91%) follow the path of 1-0-1. In fact, more than 95% of the non-general wards (i.e., 0 in the path) belong to one of the ICU-type wards. Generally, we consider these 1-0-1 patients, "GW to ICU to GW" patients. The remaining patients with paths 1-0-0 and 1-1-0 are those who initially stayed in general wards, and finally are discharged from a non-general ward. Very few patients make two transfers between three general wards (path 1-1-1).

ED-GW and EL patients with three-time transfer

Eight possible paths exist for the 905 three-time transfer patients. We aggregate some paths when displaying the statistics in Table 15. First, paths 1-0-0-0, 1-1-0-0, and 1-1-1-0 are grouped together. This group represents the patients initially admitted to a GW but discharged from a non-GW. There is no back and forth between GWs and non-GWs. Second, paths 1-0-0-1, 1-0-1-1, and 1-1-0-1 are grouped together. This group represents the patients who are initially admitted to a GW, transferred to a non-GW during the stay, and finally discharged from a GW. This group constitutes the majority of the 905 patients. Finally, the remaining two paths, 1-0-1-0 and 1-1-1-1, form their own group. Again, we can see that patients rarely make three transfers between four general wards.

Connection to the transfer-class patients in the stochastic model

The proposed stochastic model has a "single-pass" structure as stated in Section 3 of [25]. As a result, the model ignores all transfer activities inside the general wards. In other words, patients with paths 1-1, 1-1-1, 1-1-1-1 are treated the same as patients without any transfer activities. Correspondingly, patients with paths like 1-1-0 or 1-0-1-1 are treated as patients with paths 1-0 or 1-0-1, respectively.

Now we describe the connection between the transfer-class patients in the model and the real transfer patients from the empirical data. The proposed stochastic model in [25] captures two groups of real transfer patients. They are one-time transfer patients (from GWs to ICU-type wards), and two-time transfer patients (from GWs to ICU-type wards, and then back to GWs). Though these transfer activities are claimed to be between GWs and ICU-type wards, in fact, the ICU-type wards represent all the non-GWs that the model does not explicitly include. These non-GWs are referred to as ICU-type wards for convenience in the main paper. Therefore, the patients with one-time transfer cover the patients with the following paths in the data: 1-0, 1-0-0, 1-1-0, 1-0-0-0, 1-1-0-0, and 1-1-1-0. The two-time transfer patients cover the patients with the following paths in the data: 1-0-1, 1-0-0-1, 1-0-1-1, 1-1-0-1. In total, the patients captured by the transfer-class patients in the model sum up to 6670 patients, which is half of all ED-GW and EL transfer patients, and around 7%

ward	41	42	43	44	48	51	52	53	54	55	56
flow out	287	599	353	275	46	338	352	377	566	256	124
flow in	87	42	288	147	70	57	118	800	197	785	151
ward	57	57O	58	63	64	66	76	78	86	96	**total**
flow out	126	165	267	141	669	277	101	199	80	20	5618
flow in	220	458	434	398	632	298	190	156	81	9	5618

Table 16: Number of patients transferred in and transferred out for each ward considering only the 5618 patients with one-time transfer between two GWs (using combined data). Two OG wards, 48 and 96, are included because some Surgery patients overflow to them.

of the total general patients volume. The other half are right-sitting transfers within general wards, which are not modeled in [25]. See more discussion on the right-siting transfers in Section 10.2.

10.2 Right-siting transfer

We separate the 5618 ED-GW and EL patients with one-time transfer between two GWs into two groups. The first group consists of those patients who are initially admitted to the wrong wards (non-primary wards) and are later transferred to a right (primary) ward. The second group consists of the remaining patients, who are likely transfer patients from a wrong class ward to a right class ward (e.g., a subsidized patient transfers from class A to class B2). The first group comprises 3133 patients; and the second group 2485 patients. Each group constitutes about 3% of the total volume of general patients.

As mentioned, the proposed stochastic network model in [25] does not capture the transfer activities between two GWs. We believe that the non-capture does not affect the problem studied in the main paper, i.e., the impact of discharge policy on ED-GW patient's waiting time, based on two observations. First, the number of patients transferred in ("flow-in") and the number of patients transferred out ("flow-out") are more or less balanced for most wards. Table 16 shows the flow-ins and flow-outs for each ward among these 5618 patients. Most wards show a balanced flow-in and flow-out volume. Certain wards, such as 53, 55, and 63, receive more patients transferred in than the patients transferred out. For ward 53 and 63, it is because these two often receive Medicine and Cardiology patients who are medically complicated from other Medicine and Cardiology wards, respectively. Orthopedic wards (51, 52, and 54) transfer more patients than they receive, possibly because they tend to place the overflow patients back to the primary wards (recall the Orthopedic wards have high overflow proportions; see Section 5.2). Second, Figure 37, which plots the transfer-out time distribution for these 5618 patients, shows that more than 85% of the transfers occur between 2pm and 10pm, the same period when most discharges occur. This observation is consistent with NUH's policy to avoid non-urgent and unnecessary transfers unless there is a surfeit of beds.

These two observations show that (i) the occupancy level in each ward would not be affected significantly due to the balance between transfer-in and transfer-out, especially considering the transfer volume (5618 patients) is small compared to the total volume of general patients; (ii) these transfers occur in the afternoon, which has little impact on the waiting times in the morning. Therefore, our proposed model is suitable to study the inpatient operations at NUH or hospitals with similar settings given that the primary focus is to eliminate the long waiting times in the morning. Our model may not be generalized to hospitals with many transfers between general wards (e.g., hospitals which try to transfer all overflow patients back to the right wards), or to studies with a focus on internal transfers and closely related topics.

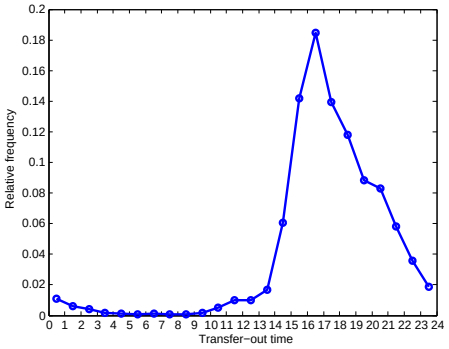


Figure 37: Transfer-out time distribution for the 5618 patients with one-time transfer between two general wards.

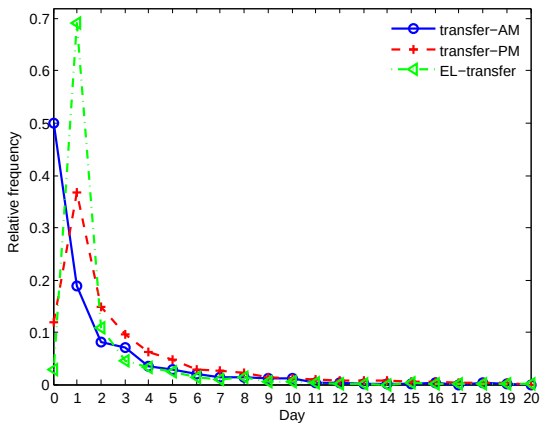
10.3 LOS distributions for one-time and two-time transfer patients in the model

As mentioned at the beginning of this section, the transfer patients captured by the stochastic model in [25] are ED-GW or EL source patients at NUH who transfer once or twice between GWs and ICU-type wards after the initial admission. For each of the *real* patients that have been modeled, her first visit to a general ward starts from the initial admission time and ends at the first transfer-out time to a ICU-type ward. If she transfers twice between GWs and ICU-type wards (a two-time transfer patient), her second visit to a general ward starts from the transfer-in time (from ICU to GW) and ends at the final discharge time.

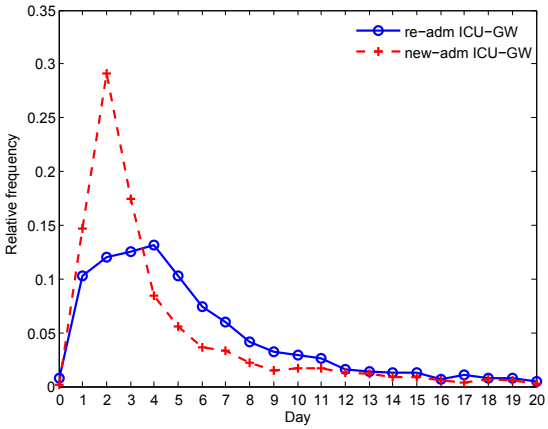
In the model, there are four classes of transfer patients. They are transfer-AM and transfer-PM patients under the ED-GW source, EL-transfer patients, and re-admitted ICU-GW patients. The first three classes of patients capture the first-visits to GWs of all the real transfer patients that have been modeled; the last class of patients are pseudo-patients, which are created to model the second-visits of those two-time transfer patients. To empirically estimate the LOS distribution for these transfer patients in the model, we use the first- and second-visit LOS of these real patients, i.e., number of nights in the corresponding visit. Specifically, we use the first-visit LOS of the real ED-GW patients with admission time before and after noon to estimate the LOS distribution for transfer-AM and transfer-PM class patients (under ED-GW source), respectively. We use the first-visit LOS of the real EL patients to estimate the LOS distribution for EL-transfer class patients. The second-visit LOS of all modeled patients who transfer twice is used to estimate the LOS distribution for re-admitted ICU-GW patients.

When empirically estimating these LOS distributions, we exclude data entries from the Orthopedic and Oncology specialties and aggregate entries from all other specialties together. We do the aggregation because (i) the empirical LOS distributions are close for patients from all specialties with the exception of Orthopedic and Oncology, and (ii) we do not have enough data points to get reliable estimation separately (for each specialty). As a result, the stochastic model in [25] assumes that the LOS distributions for the transfer-class patients do not depend on specialties. The estimated LOS distributions from the aggregated data entries is used to approximate those of the Orthopedic and Oncology patients.

We only estimate four LOS distributions, one for each of the four classes transfer patients in the model (i.e., distributions for transfer-AM, transfer-PM, EL-transfer, and re-admitted ICU-GW patients in the model). Figure 38a plots the first three empirical LOS distributions, or equivalently, the three first-stay LOS distributions of one-time real transfer patients. Figure 38b plots the LOS distribution for re-admitted ICU-GW patients, or equivalently, the second-stay LOS distribution of



(a) LOS distributions for transfer-AM and transfer-PM patients from ED-GW source, and EL-transfer patients



(b) LOS distribution for re-admitted ICU-GW patients

Figure 38: Estimated LOS distributions for transfer patients (using combined data). For references, LOS distribution for newly-admitted ICU-GW Cardiology patients is plotted in (b).

two-time real transfer patients. For comparison, in Figure 38b we also plot the LOS distribution of newly-admitted (non-transfer) ICU-GW Cardiology patients. The LOS distributions for newly-admitted ICU-GW patients are specialty dependent and are significantly different from the ones for re-admitted ICU-GW patients across all specialties.

11 Simulation model: additional details

11.1 Server pool setting and service policy

Table 17 shows the index, primary specialty, and number of servers for each server pool we use when simulating the proposed stochastic model in [25]. The table is based on the empirical study (see Table 5). Note that we adjust the number of servers in certain server pools because the proposed model cannot capture all of the constraints in bed allocation. For example, Orthopedic patients with open wounds cannot stay in the same room with patients acquired Methicillin-resistant Staphylococcus Aureus (MRSA), while the model does not differentiate between MRSA and non-MRSA patients. Thus, to replicate the specialty-level waiting time statistics, we need to reduce the number of servers in pool 7, whose primary specialty is Orthopedic.

Moreover, the model does not explicitly consider bed classes. To compensate for the inefficiency caused by class mismatch, the model assumes pools 12, 13, and 14, which correspond to the three class A/B1 wards in NUH, to be overflow pools. That is, the three pools only accept patients whose overflow trigger times are reached in the model. This adjustment is based on the facts that (i) class A/B1 wards usually do not admit class B2/C patients except for urgent situations, and (ii) the model mainly captures the performances for class B2/C patients who constitute the majority of patients at NUH. We also re-allocate some servers from the Orthopedic and Gastro-Endo pools (pools 4,7,10) into the three overflow pools, so the server numbers in these overflow pools are larger than the actual number of class A/B1 beds. This re-allocation is to capture the high overflow proportions in the Orthopedic and Gastro-Endo wards (see Section 5).

Section 4.3 of the main paper [25] discusses the stochastic model's service policy, which has four

pool ID	primary specialty	no. of servers
0	Gen Med, Respi	40
1	Gen Med, Neuro	39
2	Renal	32
3	Neuro	12
4	Gastro-Endo	38
5	Surg	41
6	Card	40
7	Ortho	49
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		557

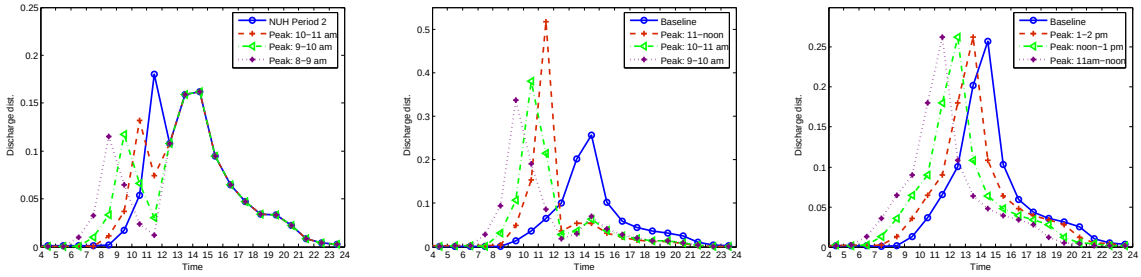
Table 17: Server pool index, primary specialty, and number of servers.

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 18: Priority of primary and overflow pools.

components: (i)picking a bed from a primary pool; (ii) picking a bed from a non-primary pool; (iii) setting an overflow trigger time; and (iv) picking a patient among a group of eligible patients. For the first two components, we construct a priority table to specify the priority of primary pools for arrivals, and the priority of non-primary pools when overflowing a patient. Table 18 is based on empirical studies in Section 5, NUH's internal bed allocation guideline [18], and discussions with NUH staff.

Models are only simplifications of the reality and cannot capture every details in the real hospital operations. Thus, the server numbers, designated specialty for the pool, and priority table we used in the simulation cannot exactly match the empirical results. To ensure the model still be relevant to the studied problem in [25], we fine-tune the model setting so that its output can match the overall hourly waiting time statistics, utilization level, specialty-level waiting times, and the overflow proportions as much as possible.



(a) 26% patients discharge by noon (b) 75% patients discharge by noon (c) Shifting the entire Period 1 discharge distribution

Figure 39: Three groups of hypothetical discharge distributions

11.2 Hypothetical discharge scenarios

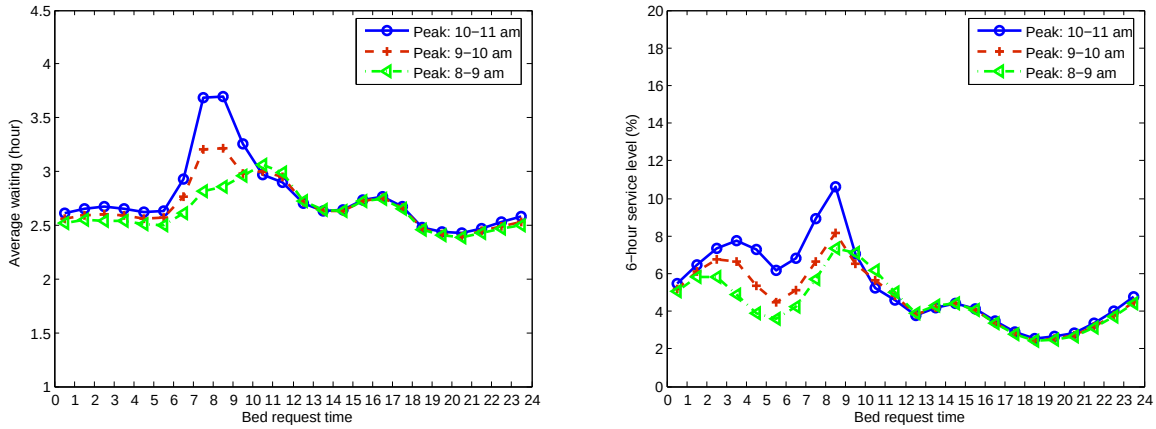
Tested discharge distributions

The simulation experiments test three groups of discharge distributions. Within each group, the discharge distribution depends on a "shift parameter". The three groups of discharge distributions are constructed as follows. Group (a) keeps the second "peak" in Period 2 discharge distribution unchanged, shifts the first peak earlier by 1, 2, and 3 hours and retains the 26% discharge before noon; (b) uses a two-peak discharge distribution similar to the one in Period 2, but 75% discharge before noon; the timing of the first peak is the same as those in (a) and in Period 2 distribution; (c) shifts the entire Period 1 discharge distribution earlier by 1, 2, and 3 hours. Among the three groups, group (a) corresponds to the scenarios with even earlier discharges; group (b) corresponds to the scenarios with more discharges by noon; and group (c) is motivated by the discharge scenarios tested in [22]. Figure 39 plots these hypothetical discharge distributions. We differentiate the distribution curves within each group by their peak time, where the peak time for groups (a) and (b) discharge distributions refer to the time of the first peak. In addition, we experiment with the "uniform" scenario for groups (a) and (b). That is, we keep 26% or 75% patients discharge before noon, and redistribute the proportions so that the discharge distribution is uniform between 8 am and noon. This scenario is also motivated by [22].

Selected simulation results

In our experiments, both the time-varying and the constant-mean allocation delay models are tested, combined with different discharge distributions as illustrated above. Figures 40 to 42 show the hourly waiting time statistics for each group of the discharge distributions. In the figures, we select several representative scenarios when the waiting time statistics can be (or almost) stabilized, i.e., combination of hypothetical early discharge policy and constant-mean allocation delay model. All simulation experiments again confirm the need for simultaneous improvement in allocation delay and discharge distribution (see Section 5.4 of [25]) in order to achieve time-stable waiting time performances.

Moreover, from the figures we observe the following. First, comparing the performances under group (a) with the first peak between 9am and 10am and the performances under group (b) with the first peak between 10am and 11am, the values are close. Recall group (a) is based on what NUH has achieved now, but shifting the first peak to an earlier time. This observation indicates that if shifting 75% patients discharge before noon is too difficult, NUH can still achieve similar



(a) Average waiting time

(b) 6-hour service level

Figure 40: Hourly waiting time statistics under the scenarios with hypothetical discharge distributions of group (a): 26% patients discharge before noon. Constant-mean allocation delay model is used.

performances by discharging those patients who are able to leave before noon as early as possible. Second, the proportion of patients discharged before noon effects performance, i.e., the waiting time is generally shorter if more patients discharge before noon. Moreover, the timing of the first peak affects waiting time. Finally, we observe that the curves under the uniform scenario of group (a) or (b) are almost identical to the curves under the scenario with peak time 9-10am in the same group.

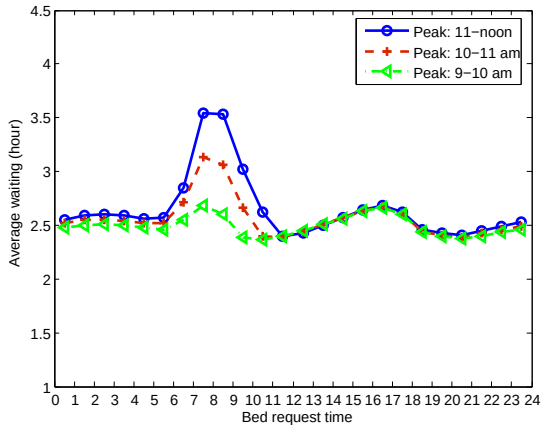
11.3 Sensitivity analysis on the choice of $p(t)$

In all the simulation experiments reported in the main paper [25], the normal allocation probability, $p(t)$, follows a step function with respect to t (see Equation (1) in Section 4.1 of [25]). In this section, we perform sensitivity analysis on the choice of $p(t)$ to study its impact on the hourly waiting time performances.

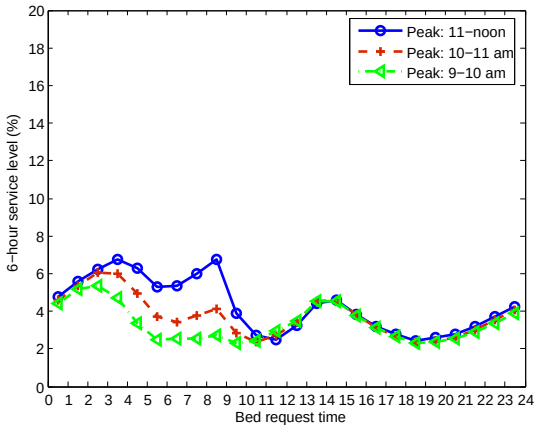
We test three settings of $p(t)$, i.e., $p(t) = 0,\ 0.5,$ or 1 for all t. Here, $p(t) = 0$ and $p(t) = 1$ serve as the lower bound and upper bound for all possible choices of $p(t)$, respectively, whereas $p(t) = 0.5$ is in-between. Figure 43 plots the hourly waiting time statistics under the baseline scenario and three new scenarios, which have the exact same settings as the baseline except the values for $p(t)$ are set to be 0, 0.5 and 1, respectively. We call the three new scenarios the *revised baseline* for each setting of $p(t)$.

Figures 44 through 46 compare each of the revised baseline with three other scenarios: (i) using Period 2 discharge distribution and the time-varying allocation delay model, i.e., Period 2 policy; (ii) using a hypothetical discharge distribution (26% discharge before, first peak between 8am and 9am) and the constant-mean allocation delay model, i.e., Period 3 policy; and (iii) reducing 10% in utilization and using the constant-mean allocation delay model. In each figure, the choice of $p(t)$ is fixed.

From these figures, we can still reach the following conclusions. First, the early discharge policy, implemented at the level that NUH achieved in Period 2, has limited impact on improving the waiting time statistics for ED-GW patients. Second, the hypothetical Period 3 policy can stabilize the hourly waiting time performances. Third, increasing capacity can reduce the daily average

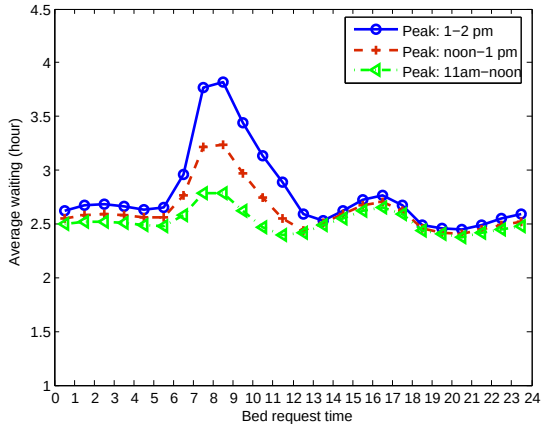


(a) Average waiting time

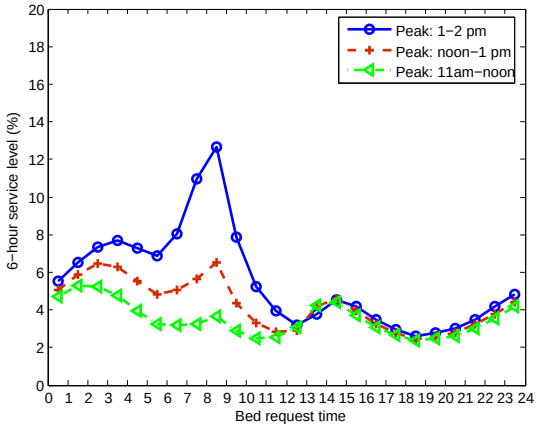


(b) 6-hour service level

Figure 41: Hourly waiting time statistics under the scenarios with hypothetical discharge distributions of group (b): 75% patients discharge before noon. Constant-mean allocation delay model is used.

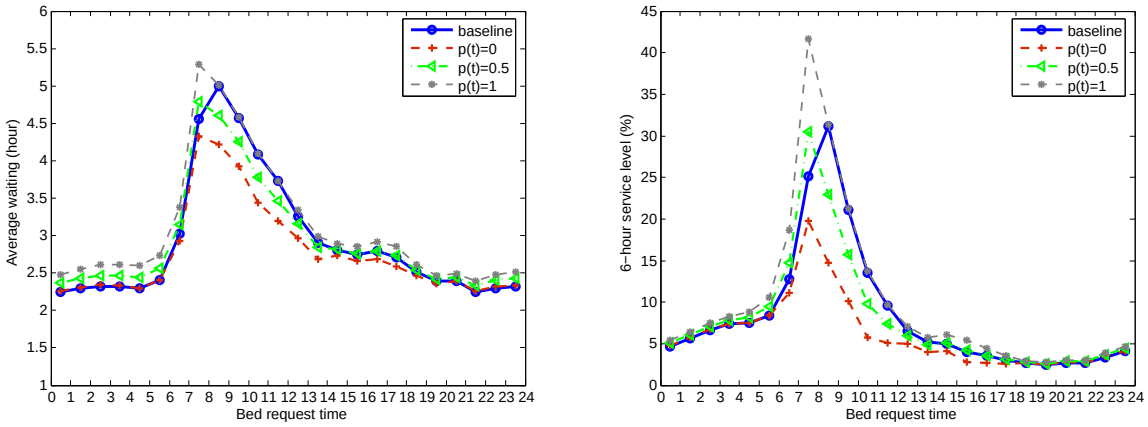


(a) Average waiting time.



(b) 6-hour service level.

Figure 42: Hourly waiting time statistics under the scenarios with hypothetical discharge distributions of group (c): shift the entire Period 1 discharge distribution. Constant-mean allocation delay model is used.



(a) Average waiting time

(b) 6-hour service level

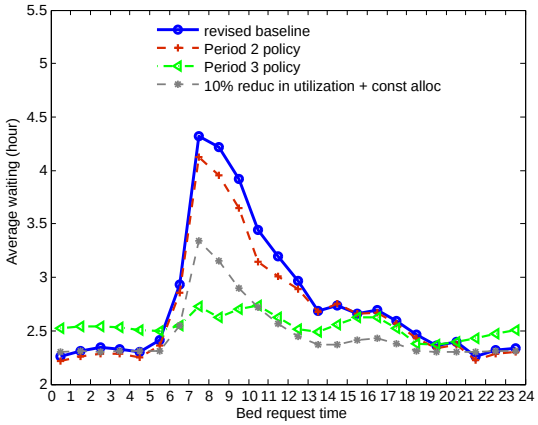
Figure 43: Hourly waiting time statistics under the baseline scenario and scenarios with different choices of $p(t)$. All simulation settings are kept the same in each scenario except the values of $p(t)$.

waiting time, but it alone cannot stabilize the hourly waiting time performances. In other words, the conclusions shown in Section 5 of [25] are not sensitive to the values for $p(t)$.

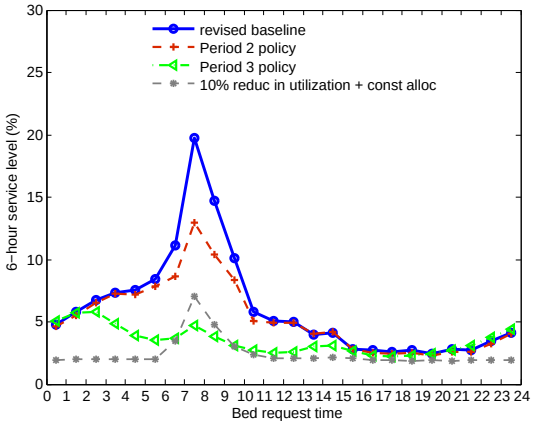
11.4 Simulation results for the overflow proportion

Overflow proportions in the simulation model are calculated in the same way as introduced in Section 5.2. The estimated overall overflow proportion from simulating the baseline scenario is 17.04%, lower than the empirical estimate of 26.95% in Period 1. Figure 18b of the main paper [25] compares the simulation estimates of overflow proportions with their empirical counterparts for each specialty. Since the simulation model does not have the concept of ward (note that server pool is motivated by but not equivalent to ward; see server pool setting in Section 11.1), we cannot compare the ward-level overflow proportions between empirical and simulation estimates.

We note that for most specialties, their overflow proportions are close between the empirical and simulation estimates. The exceptions are Surgery, General Medicine, and Neurology. The underestimation in simulation for these three specialties is the main contribution to the overall underestimation of overflow proportion across all specialties. We want to emphasize that perfectly calibrating the overflow proportions is challenging. First, it is difficult to differentiate between passive overflow and intentional overflow from empirical data. Passive overflow is triggered to avoid excessive waiting, which is the main concern in inpatient operations. Intentional overflow is usually triggered by other reasons. For example, a General Medicine patient with a potential heart problem is intentionally admitted to a Cardiology ward for telemetry care. (See similar descriptions on intentional overflow in [29].) As a result, the empirical estimation may be an overestimate of the passive overflow proportion we want to calibrate. Second, the shared server pools in our model have *complete* flexibility, i.e., each bed in such a pool can serve a patient from any primary specialty. In practice, however, complete flexibility is impossible. In fact, our empirical analysis (see Section 5.3) suggests that the shared wards have dominating preferences of serving a certain specialty regardless of the nominal bed allocation in that ward. The complete flexibility can avoid certain overflow incidences occurred in reality and leads to a smaller overflow rate in the model. For example, Neurology and General Medicine specialities share a ward (a server pool in the model). In simulation,

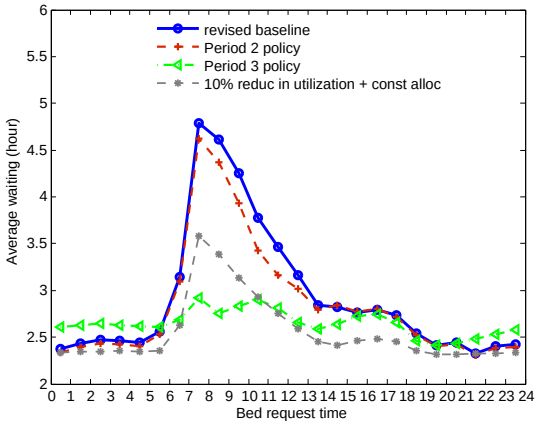


(a) Average waiting time

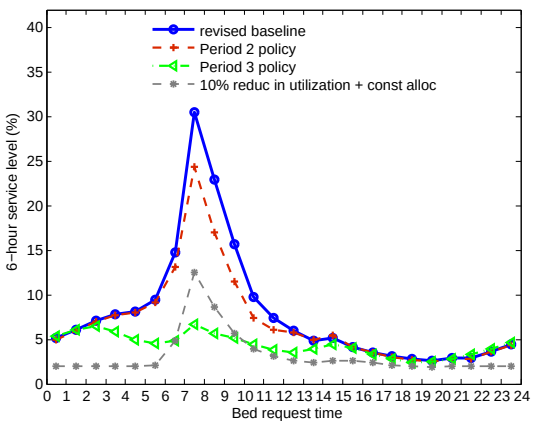


(b) 6-hour service level

Figure 44: Hourly waiting time statistics under the revised baseline scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, and (iii) 10% reduction in utilization and constant-mean allocation delay model. In all settings, $p(t) = \mathbf{0}$ for all t.

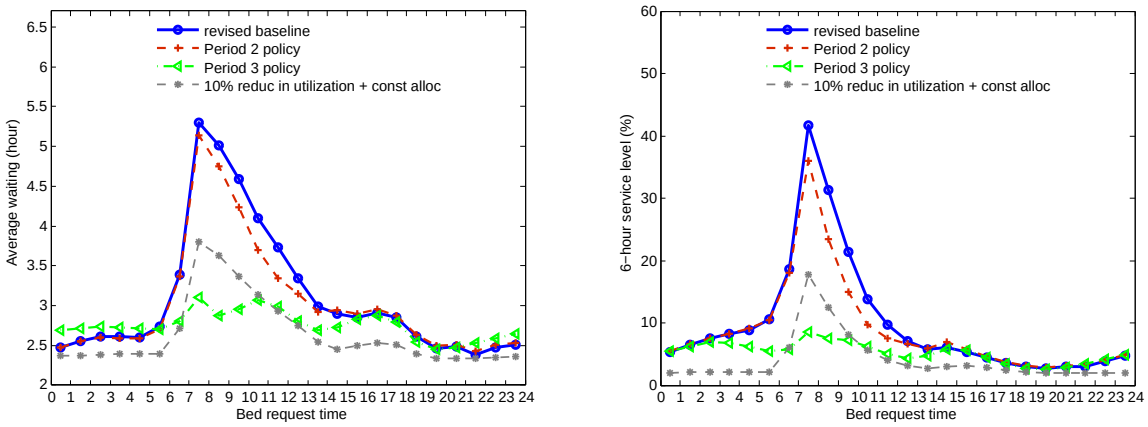


(a) Average waiting time



(b) 6-hour service level

Figure 45: Hourly waiting time statistics under the revised baseline scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, and (iii) 10% reduction in utilization and constant-mean allocation delay model. In all settings, $p(t) = \mathbf{0.5}$ for all t.



(a) Average waiting time

(b) 6-hour service level

Figure 46: Hourly waiting time statistics under the revised baseline scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, and (iii) 10% reduction in utilization and constant-mean allocation delay model. In all settings, $p(t) = \mathbf{1}$ for all t.

Neurology patients constitute 29% of all primary admissions to the shared server pool, but the proportion is only 18% for the shared ward (Ward 53) from Period 1 data. Some Neurology patients, who are admitted to the shared server pool in our model, are overflowed to other wards in practice due to a vague admission control in the shared ward. We call it "vague" since no concrete analysis can be undertaken due to the absence of data.

Section 5.2 and 5.3 of the main paper [25] demonstrate that early discharge is of little use to reduce the overall overflow proportion, even under the hypothetical Period 3 policy. We provide an intuitive explanation here. Early discharge mainly affects patients requesting beds between 7am and noon, in which period the overflow trigger time T is long ($T = 5.0$ hours from 7am to 7pm, see Section 4.2 of [25]). In the baseline scenario, for patients requesting beds between 7am and noon, primary beds are likely to become available before their waiting times exceed five hours, since most discharges start to occur from noon. Thus, few of these patients are overflowed. Given the setting of T unchanged, even though more beds become available in the morning after early discharge, the overflow proportion will be scarcely affected since there are already few overflows for morning bed-requests in the baseline scenario.

References

[1] Canadian Institute for Health Information, "Inpatient Hospitalizations and Average Length of Stay Trends in Canada, 2003-2004 and 2004-2005," 2005. [Online]. Available: https://secure.cihi.ca/free_products/hmdb_analysis_in_brief_e.pdf

[2] D. Anthony, V. K. Chetty, A. Kartha, K. McKenna, M. R. DePaoli, and B. Jack, "Re-engineering the hospital discharge: An example of a multifaceted process evaluation," in *Advances in patients safety: from research to implementation*, K. Henriksen, J. Battles, E. Marks, and D. Lewin, Eds. Rockville, MD: Agency for Healthcare Research and Quality, 2005.

[3] 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

[4] 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.

[5] 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.

[6] B. Browne and D. Kuo, "Patients admitted through the emergency department are more profitable than patients admitted electively," *Annals of Emergency Medicine*, vol. 44, no. 4, Supplement, pp. S132 –, 2004.

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

[8] 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

[9] L. V. Green, J. Soares, J. F. Giglio, and R. A. Green, "Using queueing theory to increase the effectiveness of emergency department provider staffing," *Academic Emergency Medicine*, vol. 13, no. 1, pp. 61–68, 2006.

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

[11] 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.

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

[13] P. L. Henneman, M. Lemanski, H. A. Smithline, A. Tomaszewski, and J. A. Mayforth, "Emergency department admissions are more profitable than non-emergency department admissions," *Annals of Emergency Medicine*, vol. 53, no. 2, pp. 249 – 255.e2, 2009.

[14] Hospitalist Management Advisor, "To free beds for new admissions, triage best candidates for early discharge," 2006. [Online]. Available: http://www.hcpro.com/content/62360.pdf

[15] J. Lees, J. Forsyth, T. Denison, and R. Aickin, "Valuing patients' time initiatives improve ED to ward flow," 2012. [Online]. Available: http://www.hiirc.org.nz/page/32102/valuing-patients-time-initiatives-improve/?contentType=111&section=8959

[16] S. Maman, "Uncertainty in the demand for service: The case of call centers and emergency departments," July 2009. [Online]. Available: http://ie.technion.ac.il/serveng/References/Thesis_Shimrit.pdf

[17] M. L. McCarthy, S. L. Zeger, R. Ding, S. R. Levin, J. S. Desmond, J. Lee, and D. Aronsky, "Crowding delays treatment and lengthens emergency department length of stay, even among high-acuity patients," *Ann Emerg Med*, vol. 54, no. 4, pp. 492–503 e4, 2009.

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

[19] ——, "NUH Inpatient Charges," 2012. [Online]. Available: http://www.nuh.com.sg/patients-and-visitors/appointments/hospital-charges/inpatient-charges.html

[20] NHS National Services, Scotland, "Average length of stay," 2010. [Online]. Available: http://www.isdscotland.org/Health-Topics/Hospital-Care/Inpatient-and-Day-Case-Activity/

[21] C. Osborn, *Essentials of Statistics in Health Information Technology*, 1st ed. USA: Jones and Bartlett Publishers, Inc., 2007.

[22] 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.

[23] L. Rubino, L. Stahl, and M. Chan, "Innovative approach to the aims for improvement: Emergency department patient throughput in an impacted urban setting," *The Journal of Ambulatory Care Management*, vol. 30, no. 4, pp. 327–337, 2007.

[24] O. M. Shapira, J. Gorman, C. Fitzgerald, D. Traylor, C. Hunter, H. Lazar, K. Davidson, J. Chessare, and R. Shemin, "Process improvement to smooth daily discharge time of patients after cardiac surgery improves patient flow and reduces operating room time." 2004, 10th Annual Scientific Symposium at the Institute of Healthcare Improvement. [Online]. Available: http://www.a4hi.org/symposium/2004/Shapira.pdf

[25] P. Shi, M. Chou, J. G. Dai, D. Ding, and J. Sim, "Hospital Inpatient Operations: Mathematical Models and Managerial Insights," 2012, working paper.

[26] I. Sigma Breakthrough Technologies, "Improving inpatient discharge cycle time and patient satisfaction," 2006. [Online]. Available: http://sbtionline.com/reducing-patient-discharge/

[27] Sinagpore Ministry of Health, "Waiting time for admission to ward," May 16 2012. [Online]. Available: http://www.moh.gov.sg/content/moh_web/home/statistics/healthcare_institutionstatistics/Waiting_Time_for_Admission_to_Ward.html

[28] L. Stahl, "Comprehensive emergency department and inpatient changes improve emergency department patient satisfaction, reduce bottlenecks that delay admissions," 2008. [Online]. Available: http://www.innovations.ahrq.gov/content.aspx?id=1757

[29] 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.

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

[31] VA Medical Center, Washington DC, "Getting patients home sooner: DCVAMC institutes new hospital discharge system," 2009. [Online]. Available: http://www.washingtondc.va.gov/docs/Technology-Health-Wire.pdf

[32] 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.

[33] 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.