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

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

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

J. G. Dai
School of Operations Research and Information Engineering,
Cornell University, Ithaca, NY 14853
jd694@cornell.edu

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

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

November 4, 2014

This document serves as an Online Supplement for the main paper [6]. In Section 1, we show simulation results from a more complete set of early discharge scenarios. In Section 2, we do sensitivity analysis for the baseline scenario by using alternative model settings, such as using alternative arrival processes and patient priorities. In Section 3, we evaluate the impact of early discharge and other operational policies under an increased system load. We discuss some new findings that are not observed under the load in the baseline model. Finally, in Section 4 and Section 5, we introduce additional details for the baseline simulation and provide more insights on the overflow proportions, respectively.

1 Simulation results for additional early discharge scenarios

In this section, we study the impact of early discharge on ED-GW patient's waiting time performance using a more comprehensive set of discharge distributions. In Section 1.1, we introduce the hypothetical discharge distributions that will be tested. Then in Section 1.2, we show simulation results from scenarios that use these discharge distributions. In Section 1.3, we study a scenario that uses the Period 2 early discharge distribution and includes a capacity increase at the same time. We compare the simulation output from this scenario with the empirical performance in Period 2. In Section 1.4, we demonstrate with an example that the Period 2 early discharge policy could have more significant benefits in reducing ED-GW patient's waiting time in other hospital settings.

1.1 Hypothetical discharge distributions

In our simulation experiments, we test a midnight discharge distribution and three other groups of early discharge distributions. The midnight discharge distribution simply assumes that all discharges occur at 0am each day, while the three other groups of discharge distributions are constructed as follows:

- Group (a) keeps the second discharge peak in the Period 2 discharge distribution unchanged, shifts the first discharge peak earlier by 1, 2, and 3 hours, and retains 26% discharge before noon;
- Group (b) uses a two-peak discharge distribution similar to the one in Period 2, but assumes 75% discharge before noon; the timing of the first peak is 9-10am, 10-11am, or 11am to noon;
- Group (c) shifts the entire Period 1 discharge distribution earlier by 1, 2, and 3 hours.

Figure 1 plots these three groups of discharge distributions. Note that the discharge distribution used in the Period 3 policy belongs to group (a) with the first discharge peak occurring between 8 and 9am. We differentiate the distributions within each of the three groups by their peak time, where the peak time for groups (a) and (b) refer to the time of the first discharge peak.

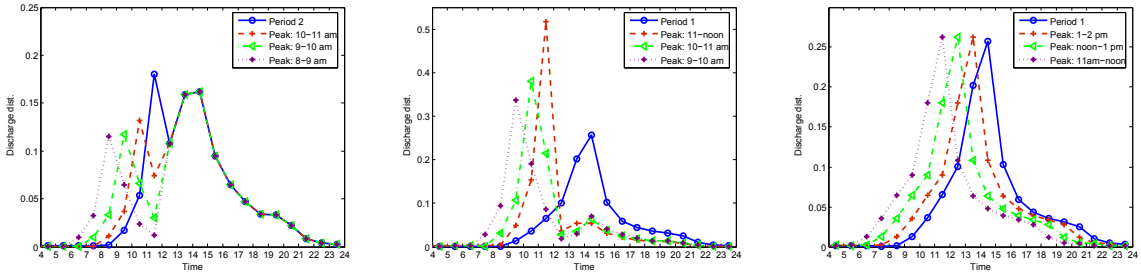
We use the midnight discharge distribution to test the maximum benefits that an early discharge policy might bring in reducing ED-GW patient's waiting time. We use groups (a) and (b) to test the impact of discharge timing and the proportion of discharge before noon on waiting time performance. Group (c) is motivated by the discharge scenarios tested in [4].

In our experiments, both the time-varying and the constant-mean allocation delay models are tested, combined with different discharge distributions as described above.

1.2 Selected simulation results

1.2.1 Simultaneous improvement is needed to flatten the waiting time curves

To achieve an approximately flattened waiting time performance, the hypothetical Period 3 policy proposed in Section 6.2 of the main paper [6] requires improvement in both the discharge timing and allocation delays. Here, we demonstrate that this simultaneous improvement is necessary. To show our results, we consider two scenarios. The first scenario uses the Period 2 discharge distribution and the constant-mean allocation delay model. The second scenario uses the same discharge distribution in the Period 3 policy and the time-varying allocation delay model. Each of these two scenarios differs from the Period 3 policy scenario only in one factor: either the discharge distribution or the allocation delay model.

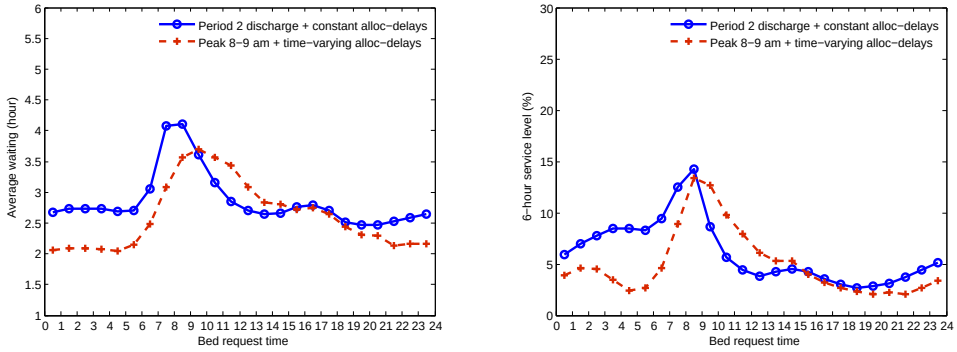


(a) 26% patients discharge by noon (b) 75% patients discharge by noon (c) Shifting the entire Period 1 discharge distribution

Figure 1: Three groups of hypothetical discharge distributions

Figure 2 plots the hourly waiting time statistics under these two scenarios. We see that in both scenarios, the average waiting time curve is not approximately flattened, i.e., the average waiting time for patients requesting beds between 7am and 11am is still about 1-2 hours longer than the daily average. The hourly 6-hour service level, though, appears to be more time-stable than the average waiting time for each scenario, especially considering the peak value is 30% in the baseline scenario.

Simulation experiments with other early discharge distributions that we have tested also confirm the need for simultaneous improvement in allocation delays and discharge timing to achieve time-stable waiting time performance.

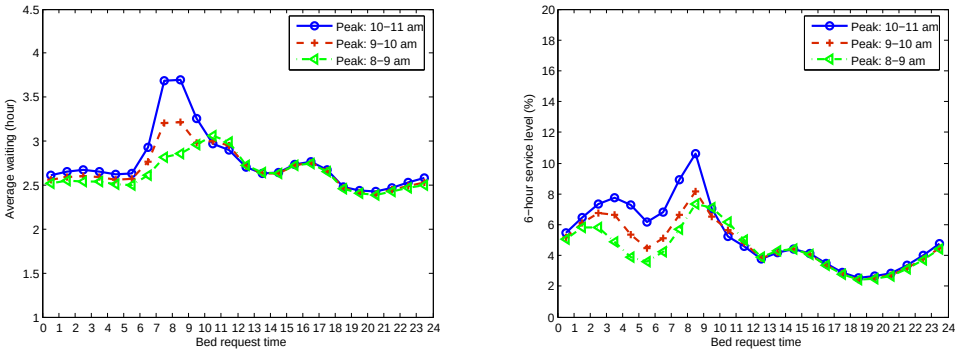


(a) Average waiting time (b) 6-hour service level

Figure 2: Hourly waiting time statistics under two scenarios. Scenario 1: Period 2 discharge distribution and constant mean allocation delays; Scenario 2: Period 3 discharge distribution and time-varying mean allocation delays.

1.2.2 Impact of the discharge timing

Figures 3 to 6 show the hourly waiting time statistics under different early discharge distributions. In each scenario, the combination of an early discharge distribution and the constant-mean allocation delay model is used; all other settings remain the same as in the baseline scenario. We observe the following from the figures.



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

Figure 3: Hourly waiting time statistics under scenarios with hypothetical discharge distributions of group (a): 26% of patients discharged before noon. A constant-mean allocation delay model is used.

First, in the National University Hospital (NUH) setting, the combination of early discharge and stabilized allocation delays can flatten the hourly waiting time performance, but has limited impact on the daily average waiting time and overflow proportions. This is true even if every patient can be discharged as early as midnight as shown in Figure 6. Indeed, in this case the daily average waiting time can only be reduced by 24 minutes from the baseline scenario, and the overflow proportion shows a less than 3% absolute reduction from the baseline scenario.

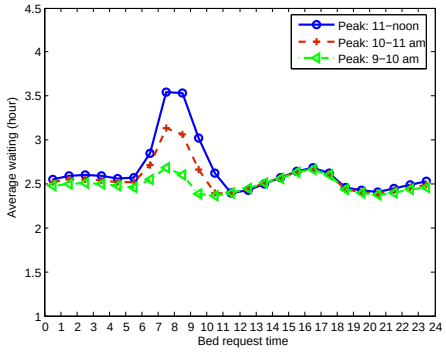
Second, the proportion of patients discharged before noon affects the waiting time performance. Generally speaking, the waiting time is shorter if more patients are discharged before noon. Moreover, we find that the timing of the first peak is important in flattening the waiting time performance. For example, if the hospital retains the first discharge peak time to occur between 11am and noon as in the Period 2 policy, even pushing 75% of the patients to be discharged before noon and stabilizing the allocation delays cannot flatten the waiting time performance.

Third, we observe that the waiting time performance under the 9-10am discharge peak scenario in group (a) is close to the performance under the 10-11am discharge peak scenario in group (b). Recall that the distributions in group (a) are based on what NUH has achieved in practice since 2010, but shift the first discharge peak to earlier time of the day. This observation indicates that if pushing 75% of the patients to be discharged before noon is too difficult, NUH (and other hospitals alike) can achieve similar waiting time performance by discharging the 26% of patients who are able to leave in the morning as early as possible.

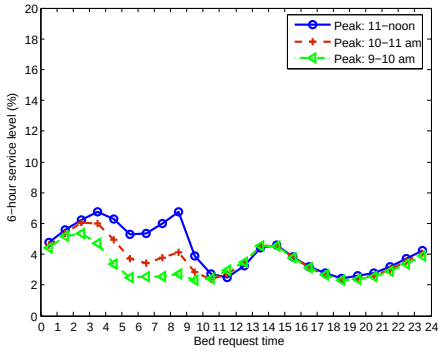
1.3 Comparing with Period 2 empirical statistics

In the introduction section of the main paper, we have discussed the changing operating environment from Period 1 to Period 2 at NUH. In Period 2, not only was the early discharge policy implemented, many other factors were also changed from Period 1. These factors include the arrival rates, the average length of stay (LOS), and the bed capacity. As a result, the bed occupancy rate (BOR) showed a 2.7% absolute reduction from Period 1 to Period 2, and the daily utilization showed a 1.7% absolute reduction. (Note that BOR and daily utilization are two slightly different concepts and are calculated in different ways; see Section 3.3 in the Companion paper [7].)

To compare with the empirical performance in Period 2, we simulate a scenario in which (i) the Period 2 discharge distribution is used, and (ii) the bed capacity is increased from the baseline scenario, producing a similar reduction in the BOR and daily utilization as we observed empirically

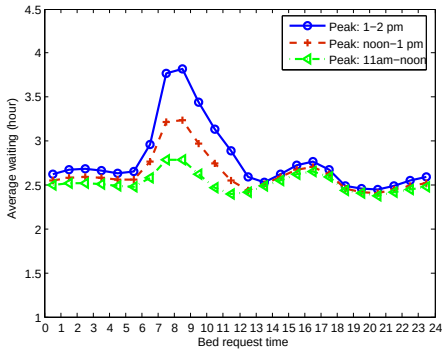


(a) Average waiting time

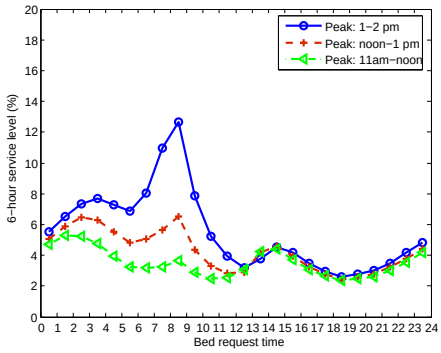


(b) 6-hour service level

Figure 4: Hourly waiting time statistics under scenarios with hypothetical discharge distributions of group (b): 75% of patients discharged before noon. A constant-mean allocation delay model is used.

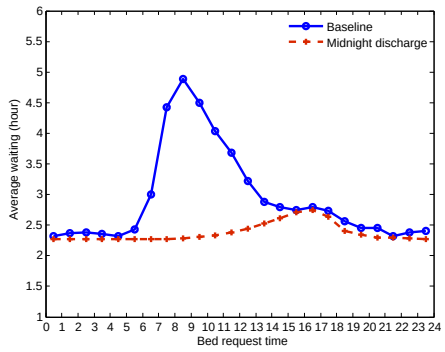


(a) Average waiting time

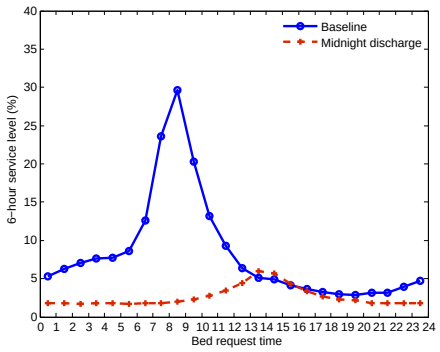


(b) 6-hour service level

Figure 5: Hourly waiting time statistics under scenarios with hypothetical discharge distributions of group (c): shift the entire Period 1 discharge distribution. A constant-mean allocation delay model is used.

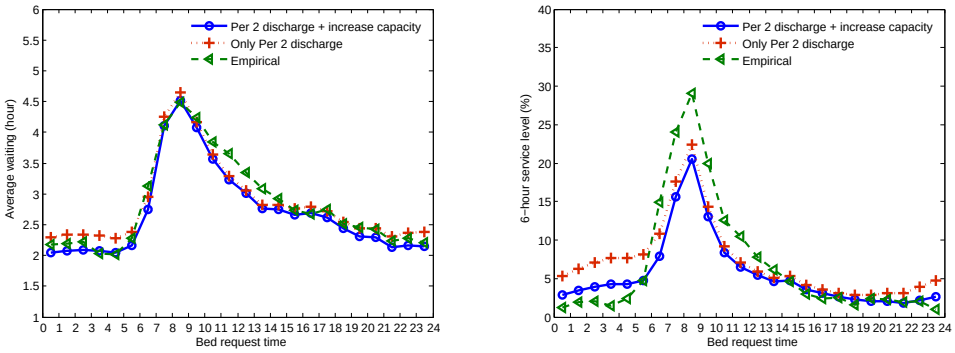


(a) Average waiting time



(b) 6-hour service level

Figure 6: Hourly waiting time statistics under the midnight discharge scenario. Constant-mean allocation delay model is used.



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

Figure 7: Simulation output compares with empirical estimates: hourly average waiting time and 6-hour service level. The empirical estimates are from using Period 2 data.

in Period 2. Other settings remain the same as in the baseline scenario. Note that this new scenario is different from the Period 2 policy scenario we introduced in Section 6.1 of the main paper, since the Period 2 policy does not include an increase in bed capacity.

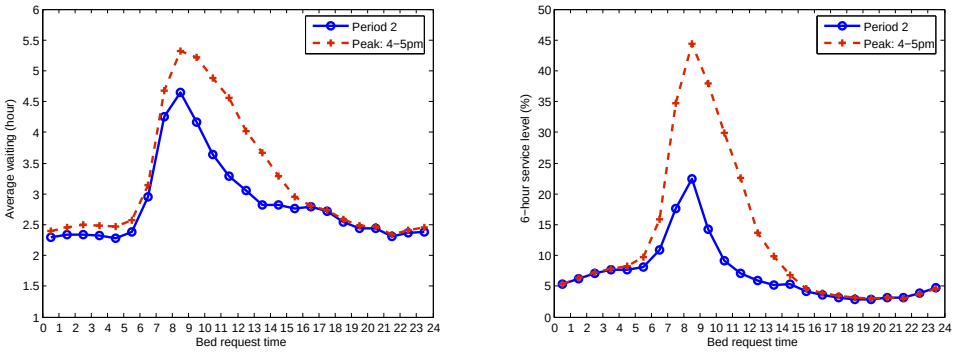
Figure 7 shows the simulation estimates of hourly waiting time statistics from the new scenario (Period 2 discharge + increasing capacity) and the empirical waiting time statistics in Period 2. For reference, we also plot the simulation estimates from the Period 2 policy scenario. From the figure, we can see that the hourly waiting time curves from the new scenario and the Period 2 policy scenario are close to the Period 2 empirical waiting time curves. In particular, the curves from the new scenario can better reproduce the empirical curves between 9pm and 6am (next day) than those from the Period 2 policy scenario.

Moreover, from Figure 1 in the main paper, we can see that the empirical hourly waiting time statistics, especially the 6-hour service level, show a reduction between 9pm and 6am in Period 2. This reduction does not appear in the simulated waiting time statistics when we change from the baseline scenario to the Period 2 policy scenario (see Figure 16 in the main paper). However, if we compare the new scenario to the baseline scenario, we observe a similar reduction in the simulated waiting time statistics between 9pm and 6am. The reason is that the new scenario includes a capacity increase, which leads to a reduction in the waiting time for patients arriving in midnight and early morning. This is also why the new scenario can better reproduce the empirical performance in Period 2, since the actual utilization in Period 2 was indeed reduced. Readers are also referred to Section 6.5 of the main paper for our discussion on how capacity increases impact waiting time statistics.

Through the observations in this section, we again see the capability of our proposed model in capturing the time-varying hourly waiting time performance and predicting the impact of various factors on the waiting time performance.

1.4 Period 2 policy could show more significant impact in other settings

In Section 7 of the main paper, we have mentioned two issues that readers should be aware of when interpreting our findings in Section 6. In particular, we want to point out here that, although the Period 2 early discharge policy shows limited impact on the waiting time statistics when compared to our baseline scenario, it does not imply this early discharge policy is not beneficial in other hospital settings. Indeed, even in Period 1, NUH manages discharge planning in a more efficient way than



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

Figure 8: Hourly waiting time statistics under the scenarios with the Period 2 discharge distribution and a hypothetical discharge distribution with the peak time at 4-5pm.

many hospitals around the world. If NUH were not discharging patients so efficiently in Period 1 (i.e., if the baseline scenario were different), we would find that implementing a Period 2 policy could bring more significant improvements to waiting time performance. We show an example below.

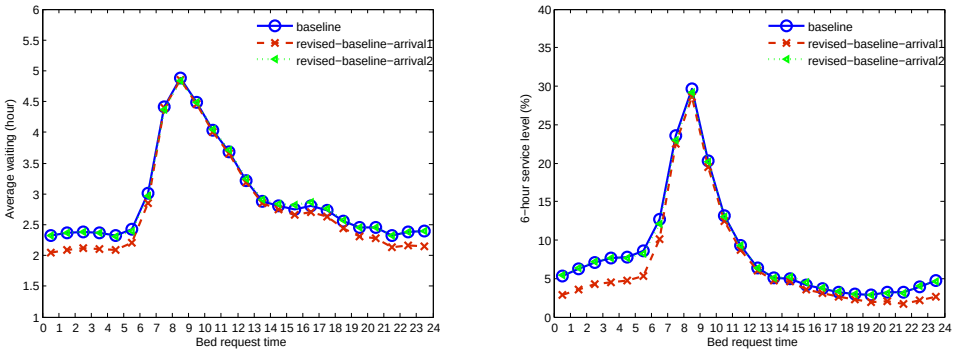
Armony et al [1] report that the discharge distribution in an Israeli hospital has a peak discharge time between 4pm and 5pm, which is two hours later than the peak discharge time in Period 1 at NUH. We now evaluate the impact of the Period 2 policy in comparison with an Israeli discharge scenario, which uses a discharge distribution similar to the one at this Israeli hospital and keeps all other settings the same as in the baseline. Figure 8 plots the hourly waiting time curves under the Israeli discharge scenario and the Period 2 policy scenario. We observe a significant improvement of waiting time statistics after implementing the Period 2 early discharge policy, even though the waiting time curves are not flattened. The daily 6-hour service level reduces from 9.26% in the Israeli discharge scenario to 5.50% in the Period 2 policy scenario (with the hourly peak value reducing from 44% to 23%). The daily average waiting time also reduces from 3.08 to 2.73 hours.

The above example indicates that implementing the Period 2 early discharge policy can be very helpful to improve waiting time performance in certain settings, especially if the hospital's current discharge timing is late. Thus, other hospitals can learn from NUH's experience in implementing the Period 2 discharge policy. The Companion paper [7] documents the details on the implementation of the Period 2 discharge policy.

2 Sensitivity analysis of different modeling settings

Sections 6.1 to 6.3 of the main paper evaluate the impact of five operations policies on waiting time performance and overflow proportions. These five policies are a Period 2 policy; a Period 3 policy; increasing bed capacity by 10%; reducing LOS by controlling the maximum stay being 14 days; and reducing the mean pre- and post-allocation delays by 30 minutes each.

To examine the robustness of the insights we have gained in Section 6 of the main paper, we test these five policies under different model settings for sensitivity analysis. These settings include using alternative arrival models (Section 2.1), changing the priority among ICU-GW, SDA and ED-GW patients (Section 2.2), using different distributions for the allocation delays (Section 2.3), and choosing different values for the normal allocation probability $p(t)$ (Section 2.4).



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

Figure 9: Hourly waiting time statistics under the baseline scenario and scenarios with different choices of arrival models. All simulation settings are kept the same in each scenario except the arrival models. In the *revised-baseline-arrival1* scenario, all four arrival processes are non-homogeneous Poisson. In the *revised-baseline-arrival2* scenario, a new batch arrival model is used for ICU-GW and SDA patients.

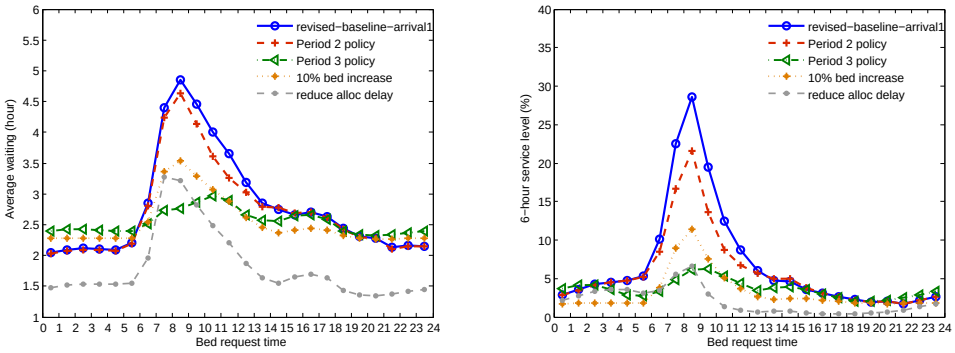
2.1 Sensitivity analysis of the arrival models

Recall that in the baseline scenario we use a time-nonhomogeneous Poisson process to model the arrivals of ED-GW patients, and non-Poisson processes to model the arrivals of other patients. (See description of the baseline setting in Sections 4.1 of the main paper [6].) Here, we perform sensitivity analysis on the choice of the arrival process models to study its impact on the hourly waiting time performance of ED-GW patients.

We test two alternative settings for the arrival processes. In the first setting, we assume the arrival processes from the four admission sources are all time-nonhomogeneous Poisson with periods of one day. The arrival rates are plotted in Figure 7 of the main paper. In the second setting, we test a modified arrival process model for ICU-GW and SDA patients based on the one proposed in Section 4.1.2 of the main paper (the arrival processes for ED-GW and EL patients remain the same as in the baseline scenario). For the modified arrival process, after we generate the A_k^j arrivals to arrive on day k from source j, we randomly assign the first arrival to a specific time of the day according to the empirical distribution of the first bed-request time. Then we assign the arrival times of the remaining $A_k^j - 1$ arrivals sequentially, 10 minutes later than the previous one. This modified arrival model is to capture a *batching* phenomenon we have observed from the bed-request times of ICU-GW and SDA patients, i.e., the inter-bed-request time is only about 10-20 minutes for most bed-requests on the same day. See additional empirical analysis in Section 6 of the Companion paper [7].

We call the scenario using the first alternative arrival setting (all non-homogeneous Poisson) the *revised-baseline-arrival1* scenario. Similarly, we call the scenario using the second alternative arrival setting (batch model for ICU-GW and SDA patients) the *revised-baseline-arrival2* scenario. Figure 9 compares the waiting time performance under the baseline scenario, the revised-baseline-arrival1 scenario, and the revised-baseline-arrival2 scenario. From the figure we can see that the waiting time performance is not sensitive to the choice of arrival models, and in particular the performance under the revised-baseline-arrival2 scenario is almost identical to that in the baseline scenario.

Next, we evaluate the five policies in comparison to the corresponding revised baseline scenario.



(a) Average waiting time

(b) 6-hour service level

Figure 10: Hourly waiting time statistics under the *revised-baseline-arrival1* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the arrival models are the same, i.e., we assume a non-homogeneous Poisson process for each admission source. For Policy (ii) to (iv), the constant-mean allocation delay model is used.

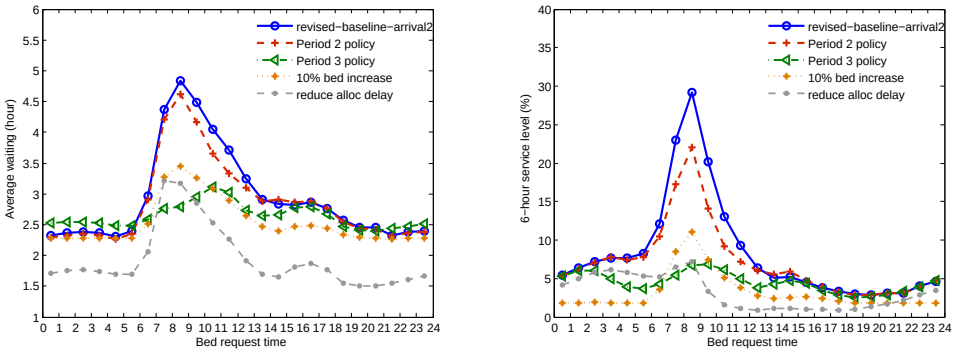
For example, to evaluate the impact of the Period 2 policy, we compare the scenario using the Period 2 discharge distribution and the first alternative arrival setting with the revised-Baseline-arrival1 scenario. All other settings not specified here remain the same as in the baseline. Figures 10 and 11 plot the hourly waiting time performance for these scenarios. Note that the performance curves under the reduced LOS scenario are almost identical to those under the increased bed capacity scenario, and we do not plot them in the figures. In each figure, the choice of the arrival model is fixed.

From these figures, we can reach the following conclusions. First, the early discharge policy, implemented at the level that NUH achieved in Period 2, has limited impact on reducing or flattening the waiting time statistics for ED-GW patients. Second, the hypothetical Period 3 policy can stabilize the hourly waiting time curves but has limited impact on the daily waiting time statistics. Third, increasing capacity, reducing LOS, or reducing mean allocation delays can reduce the daily waiting time statistics and overflow proportions, but these policies alone do not necessarily stabilize the hourly waiting time performance. In other words, the insights we gained in Section 6 of [6] are not sensitive to the choice of arrival models we have tested.

2.2 Sensitivity analysis of the patient priority

In the baseline simulation setting, EL patients have the highest priority, ED-GW patients the second, and ICU-GW and SDA patients have the lowest priority. We experiment with two alternative settings for patient priority. The first setting assigns ICU-GW and SDA patients a higher priority than ED-GW patients while keeping the highest priority for EL patients. The second setting assigns the highest priority to ICU-GW and SDA patients, followed by EL patients, and ED-GW patients have the lowest priority. We call the scenario using the first alternative priority setting the *revised-baseline-priority1* scenario. Similarly, we call the scenario using the second alternative priority setting the *revised-baseline-priority2* scenario.

Figure 12 compares the waiting time performance for ED-GW patients under the baseline scenario, the revised-baseline-priority1 scenario, and the revised-baseline-priority2 scenario. From the figure, we can see that the hourly waiting time curves under the two scenarios with alternative



(a) Average waiting time

(b) 6-hour service level

Figure 11: Hourly waiting time statistics under the *revised-baseline-arrival2* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the arrival models are the same, i.e., we assume a batch arrival model for ICU-GW and SDA patients. For Policy (ii) to (iv), the constant-mean allocation delay model is used.

priority settings are almost identical, and they are higher than the corresponding curves from the baseline scenario. This is expected since ED-GW patients have the lowest priority in the two alternative settings, and they have to wait longer than in the baseline scenario.

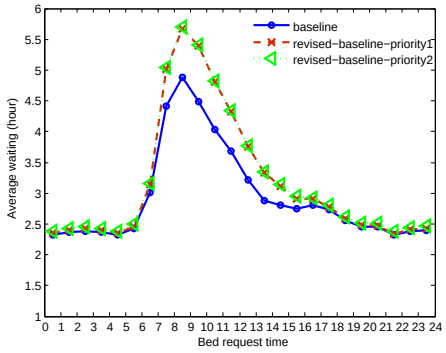
We evaluate the five policies in comparison to the corresponding revised baseline scenario. Figures 13 and 14 plot the hourly waiting time performance for these scenarios. Similar to the previous section, we do not plot the performance curves under the reduced LOS scenario since they are almost identical to those under the increased bed capacity scenario. In each figure, the priority setting is fixed.

From these figures, we can see that the insights gained in Section 6 of the main paper [6] are not sensitive to the patient priority settings that we have tested. Also note that under Period 3 policy, the hourly waiting time curves in Figures 13 and 14 are not as flattened as in the baseline, though the flattening effect is still significant. This is because ICU-GW and SDA patients, who request beds mostly in the morning, now have higher priority than ED-GW patients in the revised-baseline-priority1 and revised-baseline-priority2 scenarios. As a result, the morning congestion for ED-GW patients is more severe than in the baseline. To eliminate the excessively long waiting times for morning ED-GW bed-requests, an early discharge policy even more aggressive than Period 3 policy needs to be implemented.

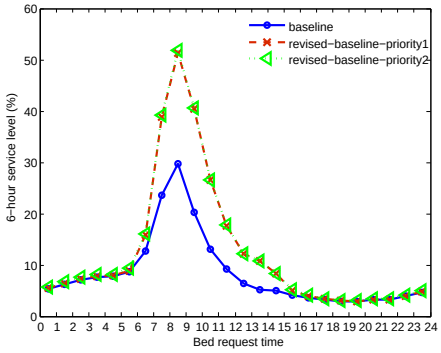
2.3 Sensitivity analysis of the allocation delay distributions

In the baseline setting, the pre- and post-allocation delays follow log-normal distributions with time-dependent means and coefficients of variation (CVs). For sensitivity analysis, we test two other distributions for the pre- and post-allocation delays: exponential and normal distributions. We assume the means (and the CVs for normal distributions) are still time-dependent, following the dashed lines with plus sign in Figure 10 of the main paper [6]. We call the scenario using the exponential allocation delay assumption the *revised-baseline-exponential* scenario. Similarly, we call the scenario using the normal allocation delay assumption the *revised-baseline-normal* scenario.

Figure 15 compares the waiting time performance for ED-GW patients under the baseline scenario, the revised-baseline-exponential scenario, and the revised-baseline-normal scenario. From the

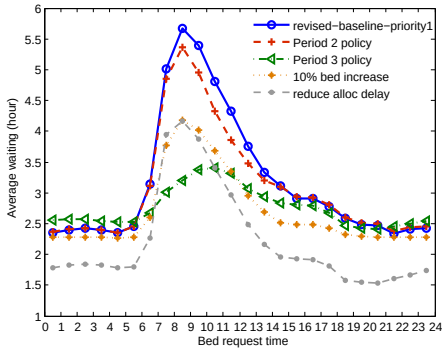


(a) Average waiting time

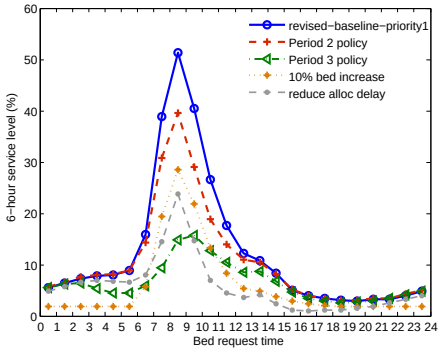


(b) 6-hour service level

Figure 12: Hourly waiting time statistics under the baseline scenario and scenarios with different patient priority settings. All simulation settings are kept the same in each scenario except patient's priority. In the *revised-baseline-priority1* scenario, EL > ICU-GW = SDA > ED-GW. In the *revised-baseline-priority2* scenario, ICU-GW = SDA > EL > ED-GW.

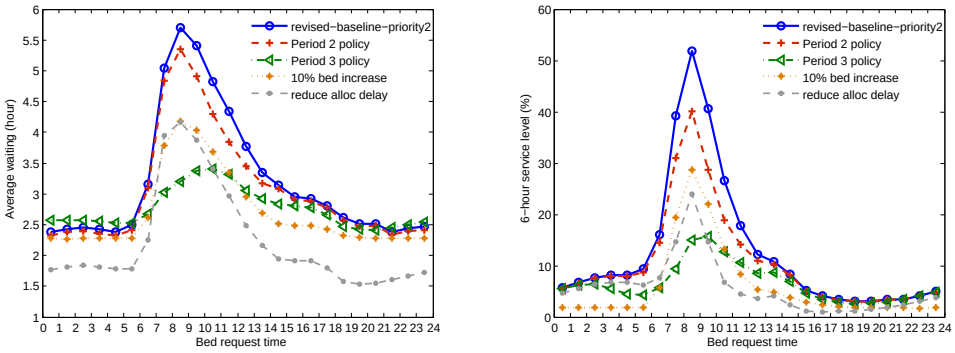


(a) Average waiting time



(b) 6-hour service level

Figure 13: Hourly waiting time statistics under the *revised-baseline-priority1* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the patient priority settings are the same (EL > ICU-GW = SDA > ED-GW). For Policy (ii) to (iv), the constant-mean allocation delay model is used.



(a) Average waiting time

(b) 6-hour service level

Figure 14: Hourly waiting time statistics under the *revised-baseline-priority2* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the patient priority settings are the same (ICU-GW = SDA > EL > ED-GW). For Policy (ii) to (iv), the constant-mean allocation delay model is used.

figures we can see that the performance measures are not very sensitive to the allocation delay distributions. In fact, the hourly average waiting time curves under the three scenarios are almost identical. This is because the average waiting time is affected by the mean allocation delays, while these mean values remain the same in all three scenarios. The differences in the allocation delay distributions are reflected through the 6-hour service level, which captures the tail distribution of the waiting times. Recall that the CV of an exponential distribution is 1, which is higher than the empirical CVs observed in Figure 10 of the main paper. Figure 15b is consistent with the common belief that higher variability contributes to longer waiting times.

We evaluate the five policies in comparison to the corresponding revised baseline scenario. Figures 16 and 17 plot the hourly waiting time performance for these scenarios. We do not plot the performance curves under the reduced LOS scenario since they are almost identical to those under the increased bed capacity scenario. In each figure, the allocation delay distributions are fixed.

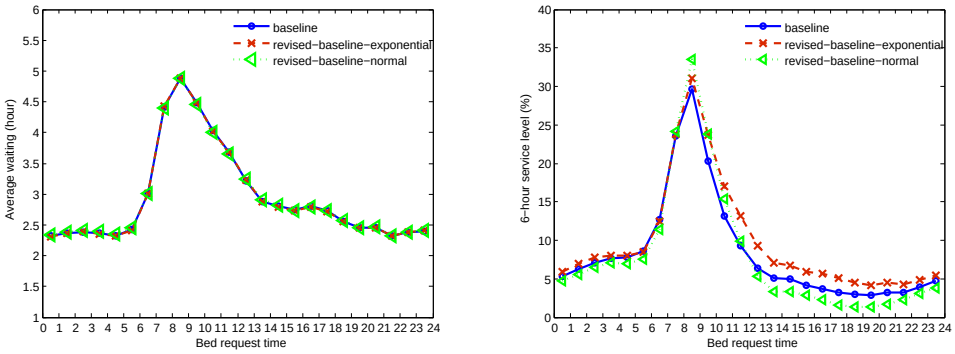
Not surprisingly, the insights gained in Section 6 of the main paper are robust with respect to the tested allocation delay distributions, since the waiting time performance is not sensitive to the tested distributions.

2.4 Sensitivity analysis of the normal allocation probability $p(t)$

In the baseline scenario, the normal allocation probability, $p(t)$ follows a step function with respect to t (see Equation (2) in Section 4.6.2 of the main paper [6]). In this section, we perform sensitivity analysis on the value of $p(t)$ to study its impact on the hourly waiting time performance. We adopt three constant functions and assume $p(t) = 0,\ 0.5,$ 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, while $p(t) = 0.5$ is in between.

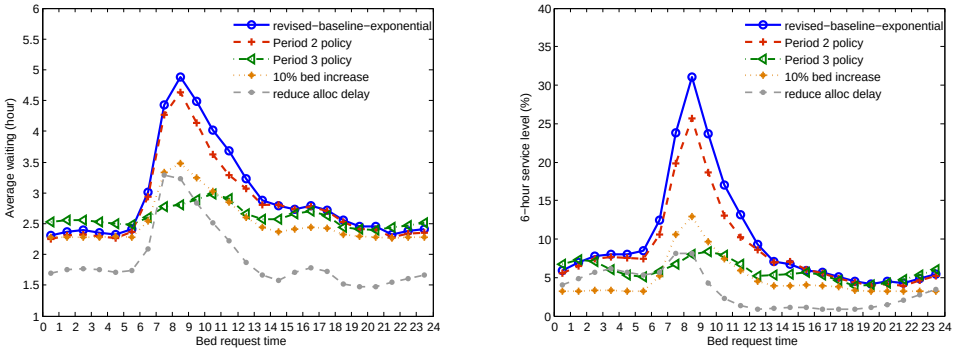
Figure 18 plots the hourly waiting time statistics under the baseline scenario and three new scenarios, which have the same settings as the baseline except for the values of $p(t)$. We call the three new scenarios the *revised-baseline-p(t)-j* scenario for $p(t) = j$ $(j = 0,\ 0.5,\ 1)$.

From Figure 18 we can see that the waiting time is longer when the value of $p(t)$ is larger, i.e., when normal-allocation mode is more frequently used than forward-allocation mode. This is because in the normal-allocation mode, the pre-allocation delay starts only after a bed becomes available,



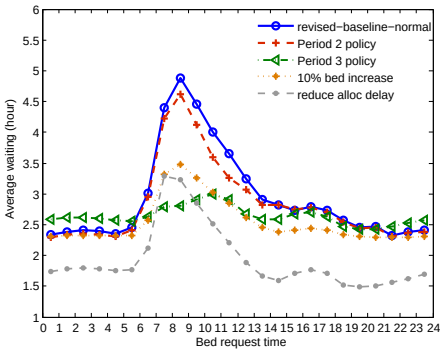
(a) Average waiting time (b) 6-hour service level

Figure 15: Hourly waiting time statistics under the baseline scenario and scenarios with different allocation delay distributions. All simulation settings are kept the same in each scenario except the distributions of allocation delays. In the *revised-baseline-exponential* scenario, exponential distributions are used for the two allocation delays. In the *revised-baseline-normal* scenario, normal distributions are used.

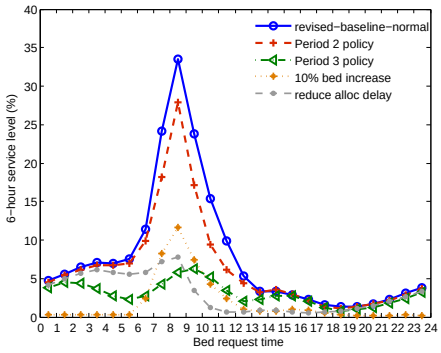


(a) Average waiting time (b) 6-hour service level

Figure 16: Hourly waiting time statistics under the *revised-baseline-exponential* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the allocation delays follow exponential distributions. For Policy (ii) to (iv), the constant-mean allocation delay model is used.

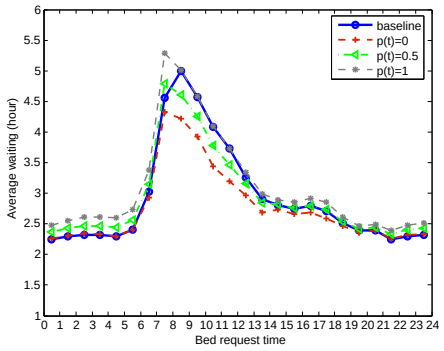


(a) Average waiting time

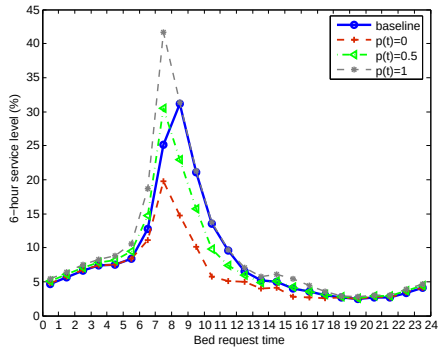


(b) 6-hour service level

Figure 17: Hourly waiting time statistics under the *revised-baseline-normal* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the allocation delays follow normal distributions. For Policy (ii) to (iv), the constant-mean allocation delay model is used.



(a) Average waiting time

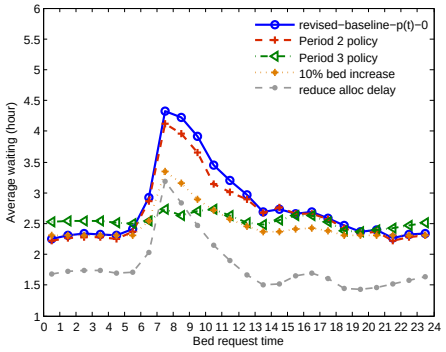


(b) 6-hour service level

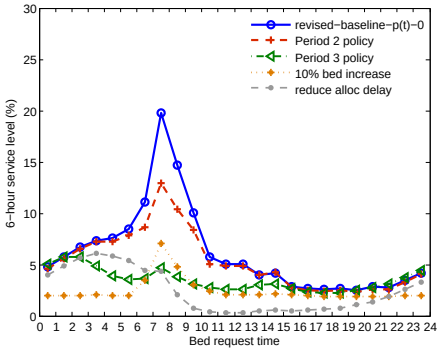
Figure 18: Hourly waiting time statistics under the baseline scenario and scenarios with different choices of $p(t)$. All simulation settings are kept the same in each scenario except the values of $p(t)$.

which is later than or the same as the bed-request time; in contrast, the pre-allocation delay always starts at the bed-request time in the forward-allocation mode. As a result, the entire waiting time for a patient in the normal-allocation mode is longer than or equal to that in the forward-allocation mode on a given sample path. Moreover, note that the value of $p(t)$ seems to have a local effect on the hourly waiting time performance. The waiting time curve for the baseline scenario coincides with one of the other three waiting time curves during certain intervals when the values of $p(t)$ are the same. For example, the baseline curve overlaps with the curve from the $p(t) = 0$ scenario between 0 and 6am since we set $p(t) = 0$ during that interval in the baseline scenario.

We evaluate the five policies in comparison to the corresponding revised baseline scenario. Figures 19 through 21 plot the hourly waiting time performance for these scenarios. We do not plot the performance curves under the reduced LOS scenario since they are almost identical to those under the increased bed capacity scenario. In each figure, the choice of $p(t)$ is fixed. Again, we can see that the insights gained in Section 6 of the main paper are not sensitive to the tested values of $p(t)$.

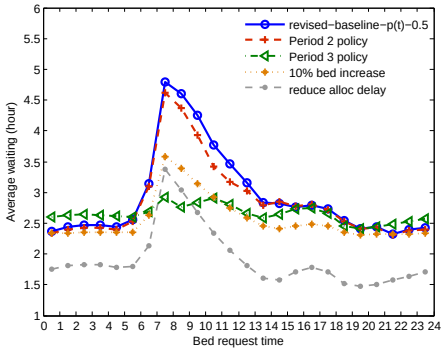


(a) Average waiting time

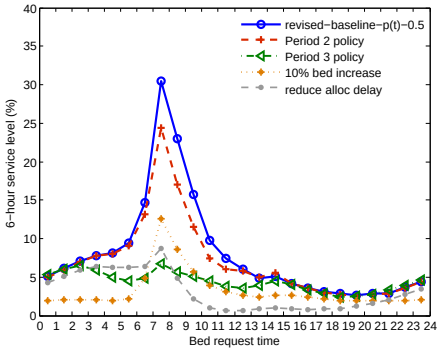


(b) 6-hour service level

Figure 19: Hourly waiting time statistics under the *revised-baseline-*$p(t)$-0 scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, $p(t) = \mathbf{0}$ for all t. For Policy (ii) to (iv), the constant-mean allocation delay model is used.

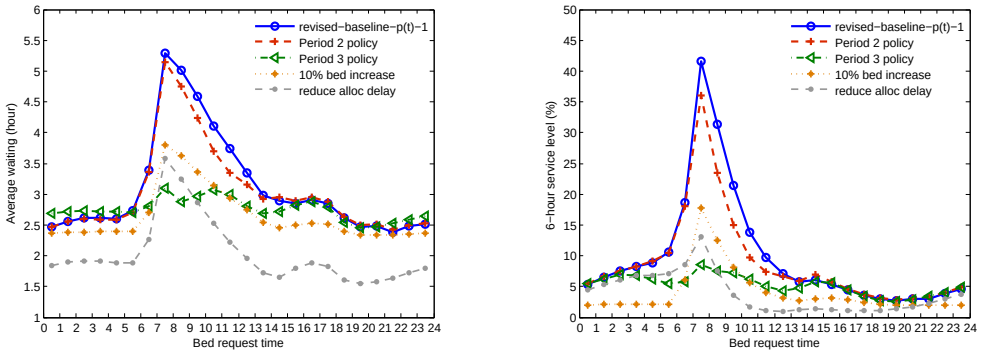


(a) Average waiting time



(b) 6-hour service level

Figure 20: Hourly waiting time statistics under the *revised-baseline-*$p(t)$-0.5 scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, $p(t) = \mathbf{0.5}$ for all t. For Policy (ii) to (iv), the constant-mean allocation delay model is used.



(a) Average waiting time

(b) 6-hour service level

Figure 21: Hourly waiting time statistics under the *revised-baseline-*$p(t)$-1 scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, $p(t) = \mathbf{1}$ for all t. For Policy (ii) to (iv), the constant-mean allocation delay model is used.

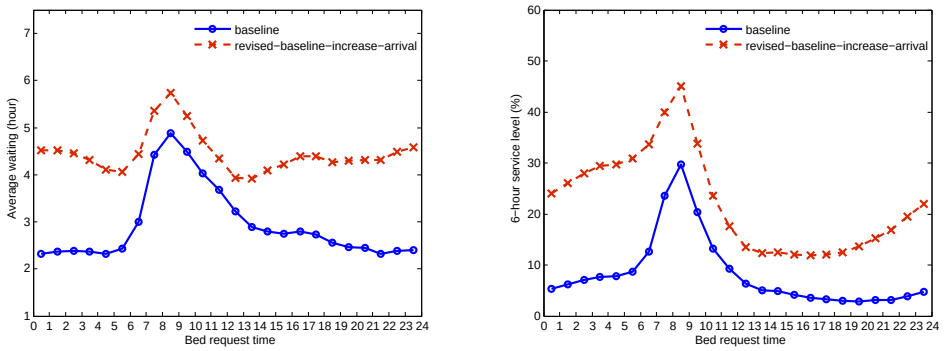
3 Sensitivity analysis of system load

In Section 6.3 of the main paper, we find that the modeled hospital queueing system is not heavily loaded in the NUH setting. The 2-3 hours average waiting time at NUH mainly comes from secondary bottlenecks such as nurse shortages rather than bed unavailability. In this section, to examine the robustness of our gained insights in a more heavily utilized setting, we increase the system load. In Section 3.1, we increase the daily arrival rate of ED-GW patients and evaluate the five operational policies that are tested in Section 2. Under the increased arrival rate setting, we find that the Period 3 policy can have a great impact on the daily waiting time statistics because of its side effect in reducing LOS, while this side effect is caused by the different LOS distributions between patients admitted before noon (AM) and after noon (PM). Thus, to separate the impact of discharge timing from the impact of reducing LOS, we eliminate the difference between the LOS distributions and re-evaluate the five policies under a similar heavily-loaded environment in Section 3.2. Finally, in Section 3.3 we summarize several conditions under which the early discharge policy can significantly impact the daily waiting time performance.

3.1 Impact of the five policies under the increased arrival rate setting

We increase the daily arrival rate of ED-GW patients by 7% from the baseline setting, similar to the increase from Period 1 to Period 2 we empirically observed. When all other settings remain the same as in the baseline, simulation shows the utilization under the increased arrival scenario becomes 93%, and the daily average waiting time and 6-hour service level become 4.37 hours and 18.60%, respectively. In other words, we create a more capacity-constrained scenario than the baseline scenario, and we call this new scenario the *revised-baseline-increase-arrival* scenario. Figure 22 compares the hourly waiting time curves between the new scenario and the baseline scenario. The curves from the new scenario have similar shapes as the curves from the baseline scenario, but are higher than the latter because of the increased system load.

We evaluate the impact of the five policies under the increased arrival rate setting and compare them with the revised-baseline-increase-arrival scenario. Figures 23 plots the hourly waiting time performance for these scenarios. Note that the performance curves under the reduced LOS scenario



(a) Average waiting time

(b) 6-hour service level

Figure 22: Hourly waiting time statistics under the baseline scenario and the scenario with increased arrival rate (*revised-baseline-increase-arrival scenario*).

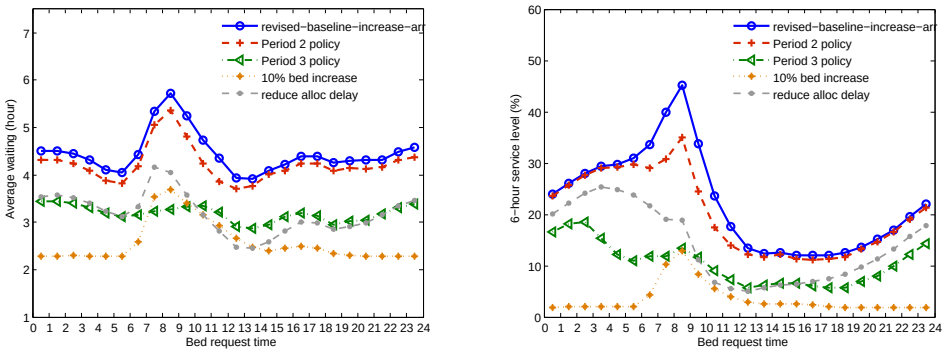
are almost identical to those under the increased bed capacity scenario, and we do not plot them in the figures.

From these figures, we can see that most conclusions we get in Section 6 of the main paper [6] still hold. First, the Period 2 early discharge policy has limited impact on reducing or flattening the waiting time statistics for ED-GW patients. Second, increasing capacity, reducing LOS, or reducing mean allocation delays can reduce the daily waiting time statistics and overflow proportions, but these policies alone cannot stabilize the hourly waiting time performance. In particular, comparing to the revised-baseline-increase-arrival scenario, increasing 10% bed capacity here reduces the daily average waiting time from 4.37 hours to 2.49 hours and the 6-hour service level from 18.60% to 2.82%, a much more significant impact on reducing the daily waiting time statistics than what we observed under the original NUH setting. This is expected because increasing capacity can greatly reduce system congestion and patient waiting time in a capacity-constrained setting, but has smaller impact if the system is not heavily loaded.

An exception is that the hypothetical Period 3 policy now not only stabilizes the hourly waiting time, but also has significant impact on the daily waiting time statistics. The daily average waiting time is reduced from 4.37 hours in the revised-baseline-increase-arrival scenario to 3.16 hours in the Period 3 policy scenario, and the 6-hour service level is reduced from 18.60% to 9.41%. The large reduction in the daily waiting times is mainly because of our assumption that the AM-admitted and PM-admitted ED-GW patients have different LOS distributions. This assumption is supported by our empirical study at NUH; see Section 4.3 of the main paper which shows that the mean LOS of AM-admitted ED-GW patients is about 1 day less than the mean LOS of PM-admitted patients across all specialties. After the Period 3 early discharge, more morning arrivals can be admitted before noon instead of waiting till the afternoon, and they become AM-admitted patients. As a result, the LOS is reduced and eventually the system utilization is reduced. We further verify this argument in the next section.

3.2 Impact of the five policies without the AM/PM difference in LOS

In this section, we assume that the AM-admitted ED-GW patients have the same LOS distributions as PM-admitted ED-GW patients for each specialty. We do so to eliminate the side effect of reducing LOS and to gain insights into the impact of discharge timing when we evaluate early discharge policies.



(a) Average waiting time

(b) 6-hour service level

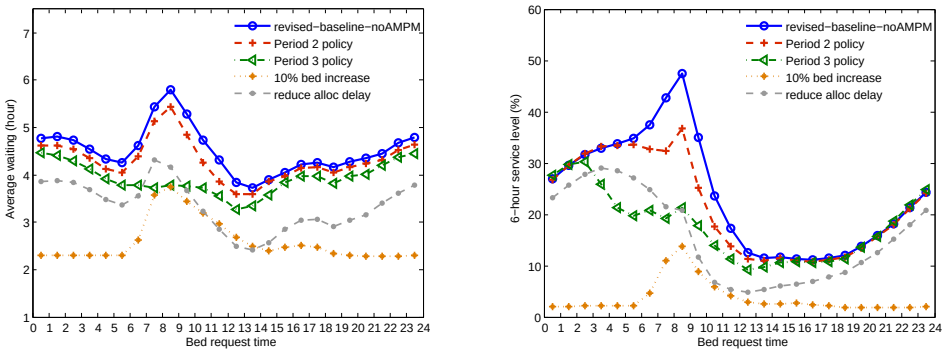
Figure 23: Hourly waiting time statistics under the *revised-baseline-increase-arrival* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the daily arrival rate for ED-GW patients is increased by 7% from the baseline setting. For Policy (ii) to (iv), the constant-mean allocation delay model is used.

Because the AM-admitted patients now have longer average LOS, we adjust the number of servers to create a capacity-constrained setting that has a similar system load as the revised-baseline-increase-arrival scenario introduced in the previous section. All other settings remain the same as in the baseline scenario. We call this scenario, without the difference in LOS distributions between AM- and PM-admitted patients (or AM/PM difference for short), the *revised-baseline-noAMPM* scenario. Simulation shows the system utilization under this new scenario is 94%. The daily average waiting time and 6-hour service level become 4.38 hours and 19.34%, respectively, which are similar to the values in the revised-baseline-increase-arrival scenario. The hourly waiting time curves under this new scenario are also close to those under the revised-baseline-increase-arrival scenario; see the solid lines in Figure 24.

We re-evaluate the impact of the five policies without the AM/PM difference in LOS. Figure 24 plots the hourly waiting time curves under these policies. Note that the performance curves under the reduced LOS scenario (i.e., control maximum LOS to be 14 days) are almost identical to those under the increased bed capacity scenario, so we do not plot them in the figure.

Comparing Figure 24 with Figure 23, we can see that Period 2 policy, increasing capacity, and reducing mean allocation delays show similar impact on the waiting time statistics no matter whether we consider the AM/PM difference in LOS or not. However, Period 3 policy shows a very different impact after we eliminate the AM/PM difference: it approximately flattens the hourly waiting time curves, but has limited impact on reducing the daily waiting time statistics. The daily average waiting time is reduced from 4.38 hours in the revised-baseline-noAMPM scenario to 3.92 hours in the Period 3 policy scenario, and the 6-hour service level is only reduced from 19.34% to 16.18%. This observation is consistent with what we get in Section 6.2 of the main paper.

In addition, we study the impact of the AM/PM difference in LOS when the system is not heavily loaded. We develop a *revised-baseline-noAMPM-normal-load* scenario by (i) assuming the AM-admitted patients have the same LOS distributions as PM-admitted patients and (ii) adjusting the number of servers to reach a similar system load as in the baseline scenario. Under this scenario, the daily average waiting time and 6-hour service level from simulation estimates are 2.80 hours and 6.27%, respectively, close to the values in the baseline scenario. We re-evaluate the impact of the



(a) Average waiting time

(b) 6-hour service level

Figure 24: Hourly waiting time statistics under the *revised-baseline-noAMPM* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the AM-admitted patients have the same LOS distributions as PM-admitted patients. For Policy (ii) to (iv), the constant-mean allocation delay model is used.

five policies under this lower system load. Figures 25 plots the hourly waiting time performance for these scenarios. Comparing the performance curves in Figures 25 to those in Figures 16 to 18 of the main paper, we can see that the five policies show similar impact with or without considering the AM/PM difference. In particular, the side effect of reducing LOS brought by the early discharge policy dost not show much impact on the waiting time when the system is not heavily loaded.

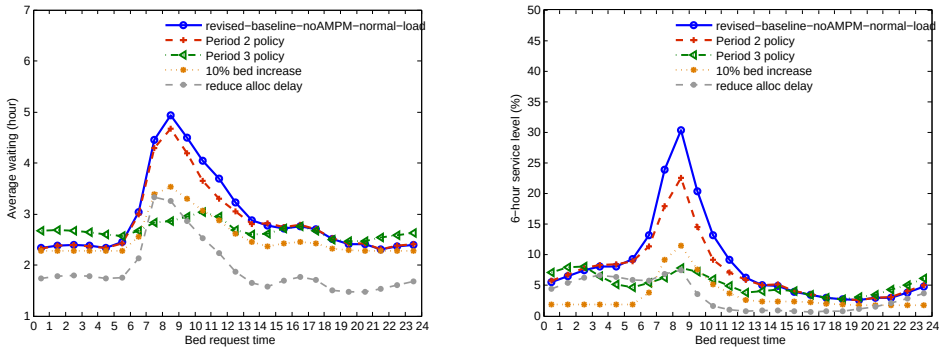
3.3 Conditions for an early discharge policy to significantly impact the daily waiting time performance

Based on our simulation findings in Sections 3.1 and 3.2, we summarize here a few conditions under which an early discharge policy can show a significant impact on the daily waiting time statistics.

First, when the LOS of AM-admitted patients is shorter than the LOS of PM-admitted patients, implementing an early discharge policy can reduce LOS in addition to shifting the discharge timing. Therefore, early discharge can significantly affect the system load and reduce the daily waiting time statistics. However, when the AM- and PM-admitted patients have the same LOS distributions, the early discharge policy no longer affects the LOS but only influences the discharge timing. In this case, shifting the discharge timing can flatten the waiting time curve but has limited impact on reducing the daily waiting time statistics. In Section 6.5 of the main paper, we have provided some intuitive explanation for why reducing LOS and shifting discharge timing have different impacts on the daily and hourly waiting time performance.

Second, given that the early discharge policy shows a side effect of reducing LOS, the impact of reducing LOS on the waiting time statistics is significant only when the system is heavily loaded. In the NUH setting, the Period 3 policy shows a limited impact on the daily performance even if we use different LOS distributions for AM- and PM-admitted patients (see Section 6.2 of the main paper). The main reason is that the system load is not high enough in the NUH setting.

Third, in order for an early discharge policy to show a significant side effect of reducing LOS, the discharge timing needs to be early enough. Unlike the Period 3 policy, the Period 2 early discharge policy cannot reduce the daily waiting time statistics much, no matter whether we differentiate between AM- and PM-admitted patients or not. This is because, under the Period 2 policy, the first discharge peak is between 11am and noon. Even after implementing the Period 2 early discharge,



(a) Average waiting time

(b) 6-hour service level

Figure 25: Hourly waiting time statistics under the *revised-baseline-noAMPM-normal-load* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the AM-admitted patients have the same LOS distributions as PM-admitted patients. For Policy (ii) to (iv), the constant-mean allocation delay model is used.

most morning arrivals still have to be admitted after noon due to the allocation delays (which on average takes about 2 hours) and the LOS is not effectively reduced.

Finally, we want to point out the need of future research to identify the factors causing the AM/PM difference in LOS. This line of research can help us better understand whether the 1-day difference in the mean LOS between AM- and PM-admitted patients will still exist when more patients are admitted in the morning than what we observed so far. Eventually, this research can help us generate more comprehensive insights into the benefits of early discharge policies and other operational policies.

4 Additional details for the NUH model and simulation settings

This section contains supplementary details for the baseline NUH model and simulation experiment settings. In Section 4.1, we discuss the adjustments we have made on the server pool setting. In Section 4.2, we discuss how we choose the values for the normal-allocation probability $p(t)$ in the baseline scenario. In Section 4.3, we show that our choices of the warm-up period and the length and number of batches are appropriate for our simulation study.

4.1 Server pool setting in the NUH model

Table 1 of the main paper lists the number of servers and primary specialties for each of the 15 server pools in the baseline NUH model. These numbers of servers and primary specialties are determined in two steps. Recall that each pool in the model corresponds to a ward or a set of similar wards at NUH. In the first step, we set up the primary specialty for each pool based on this correspondence and choose the initial number of servers using NUH's bed capacity at the *end* of Period 1. Then, we make several adjustments to obtain the final setting listed in Table 1. Below, we specify these adjustments.

The first adjustment we have made is to assume that pools 12, 13, and 14 are three overflow server pools. In the model, these three pools only accept patients who have waited beyond the overflow

trigger time. Note that these patients may belong to any medical specialties. Thus, these three pools no longer have primary specialties. As to be explained below, the three pools correspond to the three wards at NUH that have Class A or B1 beds, and we obtain the initial number of servers for pools 12-14 based on this relationship. In this adjustment we do not change the number of servers; but we do change how pools 12-14 function in the model. Note that we still need to differentiate the three overflow pools since for different specialties, the priorities when choosing among the three pools are different; see Table 2 in the main paper.

The rationale of using the three overflow server pools is that Class A/B1 wards at NUH operate differently from other wards. At NUH, beds have different classes: Patients staying at Class A/B1 beds share a room with 0-3 other patients; staying at these beds are expensive. Patients staying at Class B2 or C beds have to share a room with 5-7 other patients; staying at these beds are much cheaper due to the heavy subsidy from government for these two classes of beds (see [3] for the current price list). The majority of the patients at NUH intend to stay in a Class B2 or C bed. In general, they do not receive a free upgrade to Class A/B1 beds even if there is no Class B2/C bed available. Only when Class B2/C beds are in severe shortage, NUH may overflow (upgrade) patients to Class A/B1 beds to avoid extremely long wait. Based on this practice, we create the three overflow server pools to capture such "admission control" phenomenon. This adjustment can also partly compensate for the deficiency that we are currently not able to explicitly model bed classes due to data unavailability. Later in Section 5.2, we will see that if we do not make this adjustment but assume Class A/B1 beds can admit any patients immediately without any wait, the waiting time performance and overflow proportion cannot match the empirical estimates.

The second adjustment is that we re-allocate some servers from the Orthopedic, Renal, Gastro-Endo, and Orthopedic/Surgery pools (pools 2, 4, 7, 10) into the three overflow server pools. Thus, the number of servers in the overflow pools are larger than the actual number of Class A/B1 beds, and the number of servers in pools 2, 4, 7, 10 are less than the actual capacity. This re-allocation is to capture the unusually high overflow proportion in the Orthopedic, Renal and Gastro-Endo wards; see empirical evidence in Section 2.3 (Figure 8) of [5]. According to the NUH staff, one possible reason for such high overflow proportion is that the hospital tends to "reserve" some capacity in these wards, so that when a patient waits (or is expected to wait) too long, he or she can be overflowed to the reserved capacity. In a way, this practice is similar to the admission control phenomenon we discussed above for Class A/B1 wards. Therefore, we re-allocate these capacities to the three overflow pools in the model.

Note that there is no data for us to directly estimate how many beds we should re-allocate from one pool to another. The final setting in Table 1 is obtained through trial-and-error experiments so that the waiting time statistics from simulation are close to the empirical estimates for each specialty.

The last adjustment we have made is to reduce the number of servers in certain pools. The reason is that the bed capacity at NUH had increased in Period 1, while our basis to choose the initial number of servers for each pool is the bed capacity at the end of Period 1. Thus, these initial numbers overestimate the average capacity during the entire Period 1. The major reductions in the number of servers are done for Cardiology and Respiratory pools (pool 6, 9, and 11), because the actual increase in the bed capacity for these two specialties was the largest during Period 1. Minor reductions in the number of servers are done for some other pools. Again, there is no detailed data for us to directly estimate the reduction, and we have used simulation experiments to determine the final pool setting.

4.2 The normal-allocation probability $p(t)$

Recall from Section 3.4 in the main paper that, when a patient makes a bed-request at time t and there is no primary bed available at the time, we assume that with probability $p(t)$ the allocation mode for the patient is the normal-allocation mode, meaning this patient will wait until a bed is available before starting to experience the pre-allocation delay. In this section, we explain the rationale for using $p(t)$ given in Equation(2) of the main paper. Consistent with the notation in the main paper, we use $h(t) = 24 \cdot (t - \text{floor}(t))$ to denote the time-of-day for the bed-request time t. We use hour as the time unit for $h(t)$, and day for t.

First, the choice of $p(t) = 1$ for $h(t)$ between 8am and 12 noon is consistent with the current practice at NUH. In order to do a forward allocation, the planned discharge information should be available. Most wards do the morning rounds at about 9-11am, and nurses would only know which patients will be discharged after finishing the rounds. Thus, BMU typically receives the planned discharge information when the time is close to noon.

Second, for $h(t)$ between 2pm and 8pm, we use $\hat{p}(i)$ to empirically estimate $p(t)$ for each hour i between 2pm and 8pm. For all ED-GW patients (in the NUH data) whose bed-request time falls within hour i, we define $\hat{p}(i)$ as

$$\hat{p}(i) = \frac{\#\text{ of patients whose allocation-completion time } > \text{ bed-available time}}{\#\text{ of patients whose bed is not available at bed-request time}}. \tag{1}$$

Here, the denominator consists of the patients whose allocated bed is not available at the bed-request time, i.e., these patients have waited for a bed to become available after their bed requests. The patients included in the numerator correspond approximately to normal allocations in the model. Based on the empirical estimates, we set $p(t) = .5$ between 2pm and 8pm. Section 9.2 of the Companion paper [7] explains why patients in the numerator correspond to normal allocations and why we can use (1) to estimate $p(t)$ in certain time intervals.

Third, our empirical analysis also shows that, between 8pm and 6am the next day, there are very few (fewer than 15 each hour) normal allocations, suggesting $p(t)$ is close to zero. Therefore, we set $p(t) = 0$ for $h(t)$ between 8pm and 6am the next day.

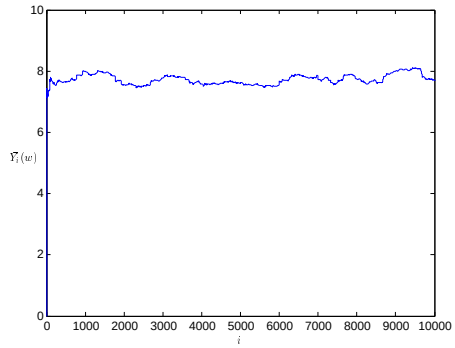
Fourth, during each of the remaining time intervals of a day, $(6, 8]$ or $(12, 2]$, we estimate $p(t)$ by interpolating its values in the neighboring intervals to avoid sudden changes of $p(t)$. The actual values of $p(t)$ in these two intervals are obtained by trial-and-error so that the simulation estimates can approximately replicate the empirical waiting time performance.

We realize that, despite our best efforts, our choice of $p(t)$ is still ad hoc. Therefore, we have conducted a sensitivity analysis of the choice of $p(t)$ in Section 2.4.

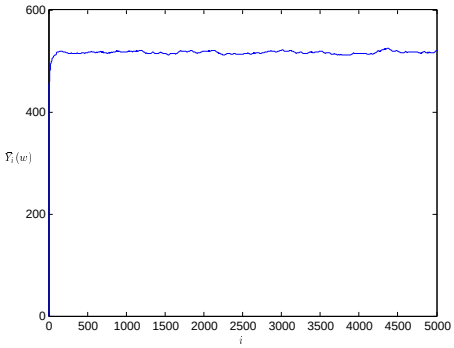
4.3 The warm-up period and the length and number of batches

In each simulation experiment, we simulate for a total of 10^6 days, and divide the simulation output into 10 batches. The performance measures are calculated by averaging the last 9 batches, with the first batch discarded to eliminate transient effects. This simulation setting is justified as below.

First, we follow the standard procedures in the literature to observe the moving average plot [2] and determine the warm-up period. We run $n = 5$ replications of the baseline scenario (5-10 replications are recommended choices in the literature), with each replication running for $m = 10^5$ days. We define Y_{ji} be the ith observation from the jth replication, where the ith observation can be a chosen performance measure on day i, e.g., the daily average buffer size or average number of busy servers on day i. Let $\bar{Y}_i = \frac{1}{n}\sum_{j=1}^{n} Y_{ji}$. For a given time-window w, we define the moving



(a) Daily average buffer size with $w = 500$



(b) Daily average number of busy servers with $w = 100$

Figure 26: Moving average plots from 5 replications. Each replication contains 10^5 days.

average as

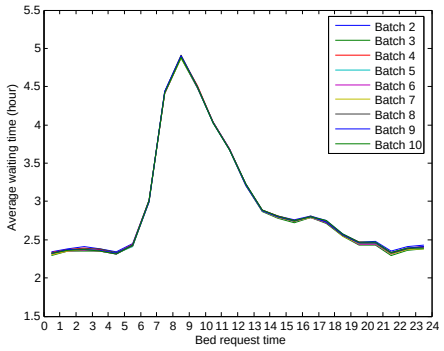
$$\bar{Y}_i(w) = \begin{cases} \frac{1}{2w+1} \sum_{s=-w}^{w} \bar{Y}_{i+s}, & \text{if } i = w+1, \ldots, m-w; \\ \frac{1}{2i-1} \sum_{s=-(i-1)}^{i-1} \bar{Y}_{i+s}, & \text{if } i = 1, \ldots, w. \end{cases}$$

The time-window w is chosen through experiments, so that we can both observe the initial transient effects and have a reasonably smooth plot after the system converges to steady state. Figure 26 shows the moving average plots for the daily average buffer size and average number of busy servers. We chose $w = 500$ for the buffer size plot and $w = 100$ for the busy server plot. It is clear that before 10^4 days, the sequence of $\{\bar{Y}_i(w)\}$ appears to have converged, indicating our choice of 10^5 days as the warm-up period is more than enough.

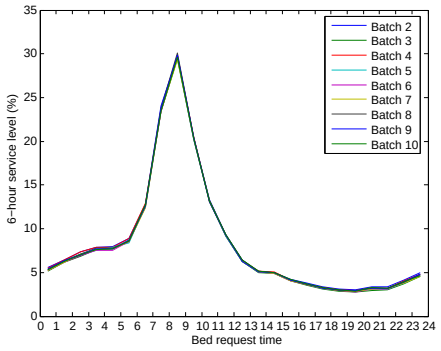
Second, Figure 27 compares the hourly waiting time statistics from each of the last nine batches under the baseline scenario (the warm-up batch is excluded), and each batch contains a total of 10^5 days. It is clear from the figure that the waiting time curves from the 9 batches are very close to each other, suggesting (i) the system is running in the steady state; and (ii) the choice of the number of batches and length of each batch is appropriate to produce a tight confidence interval for the reported batch means. Table 1 below reports the the confidence intervals for the batch means of the hourly average waiting time and the hourly 6-hour service level. For comparison, the confidence intervals of the corresponding empirical hourly waiting time statistics are also reported in the table. Table 3 in the main paper reports the confidence intervals for the batch means of the daily and specialty-level average waiting times and 6-hour service levels.

5 Additional discussion on the overflow proportion

Besides ED-GW patient's waiting time, the overflow proportion is one of the performance measures of interest to us and is monitored closely at NUH. This section provides more discussions on the overflow proportions. In Section 5.1, we discuss a few factors that have contributed to the challenge of reproducing the empirical overflow proportions with our proposed model. These factors include the server pool setting and the simplified overflow policy. To ensure that our insights are robust in these two factors, in Sections 5.2 and 5.3, we conduct sensitivity analyses under alternative server-pool settings and overflow policies. Finally, in Section 5.4, we provide some intuition on why early discharge policies show limited impact on overflow proportions as seen in Section 6 of the main paper.



(a) Average waiting time



(b) 6-hour service level

Figure 27: Hourly waiting time statistics from each batch.

	average waiting time (hour)		6-hour service level (%)	
hour	simulation	empirical	simulation	empirical
1	2.32 (2.31, 2.33)	2.41 (2.30, 2.51)	5.33 (5.25, 5.41)	3.62 (2.69, 4.56)
2	2.36 (2.36, 2.37)	2.32 (2.20, 2.43)	6.28 (6.20, 6.35)	3.54 (2.51, 4.56)
3	2.38 (2.36, 2.39)	2.31 (2.18, 2.43)	7.08 (6.97, 7.19)	3.70 (2.52, 4.89)
4	2.36 (2.35, 2.37)	2.17 (2.05, 2.28)	7.68 (7.60, 7.76)	3.71 (2.45, 4.97)
5	2.32 (2.32, 2.33)	2.19 (2.04, 2.34)	7.76 (7.68, 7.83)	4.81 (3.16, 6.47)
6	2.43 (2.42, 2.43)	2.51 (2.32, 2.70)	8.64 (8.53, 8.75)	8.70 (6.46, 10.94)
7	3.01 (3.00, 3.01)	3.24 (3.01, 3.47)	12.64 (12.54, 12.75)	18.97 (15.83, 22.10)
8	4.42 (4.41, 4.43)	4.33 (4.08, 4.57)	23.60 (23.46, 23.74)	33.21 (29.22, 37.20)
9	4.89 (4.88, 4.90)	4.64 (4.41, 4.86)	29.67 (29.52, 29.81)	36.52 (32.59, 40.46)
10	4.49 (4.49, 4.50)	4.74 (4.58, 4.89)	20.32 (20.24, 20.40)	30.84 (27.75, 33.94)
11	4.03 (4.03, 4.04)	4.22 (4.11, 4.33)	13.15 (13.10, 13.21)	16.61 (14.55, 18.68)
12	3.68 (3.68, 3.68)	3.81 (3.72, 3.90)	9.25 (9.21, 9.28)	10.99 (9.48, 12.50)
13	3.22 (3.21, 3.22)	3.37 (3.30, 3.45)	6.33 (6.29, 6.38)	7.50 (6.33, 8.67)
14	2.88 (2.88, 2.89)	3.15 (3.07, 3.22)	5.09 (5.06, 5.12)	7.24 (6.16, 8.33)
15	2.80 (2.79, 2.81)	2.88 (2.81, 2.94)	4.95 (4.91, 4.99)	5.46 (4.51, 6.40)
16	2.74 (2.74, 2.75)	2.73 (2.67, 2.80)	4.11 (4.07, 4.15)	3.87 (3.09, 4.65)
17	2.80 (2.79, 2.80)	2.65 (2.59, 2.71)	3.65 (3.61, 3.68)	2.56 (1.90, 3.22)
18	2.73 (2.72, 2.74)	2.74 (2.68, 2.80)	3.22 (3.18, 3.27)	2.56 (1.90, 3.22)
19	2.56 (2.55, 2.57)	2.48 (2.42, 2.55)	2.98 (2.93, 3.02)	1.84 (1.25, 2.42)
20	2.46 (2.45, 2.47)	2.40 (2.33, 2.48)	2.88 (2.82, 2.93)	2.25 (1.55, 2.94)
21	2.45 (2.44, 2.46)	2.32 (2.25, 2.40)	3.18 (3.11, 3.25)	1.78 (1.17, 2.39)
22	2.32 (2.31, 2.33)	2.30 (2.22, 2.38)	3.18 (3.11, 3.25)	2.59 (1.89, 3.29)
23	2.38 (2.37, 2.39)	2.33 (2.23, 2.43)	3.92 (3.83, 4.00)	3.50 (2.66, 4.34)
24	2.40 (2.39, 2.41)	2.27 (2.18, 2.36)	4.73 (4.64, 4.82)	3.13 (2.33, 3.93)

Table 1: Simulation and empirical estimates of the waiting time statistics for ED-GW patients requesting beds in each hour of the day. The simulation estimates are from simulating the baseline scenario, and the empirical estimates are from Period 1 data. The numbers in the parentheses are for the 95% confidence interval of the corresponding value. The confidence intervals for the simulation output are calculated following the batch mean method [2]; the confidence intervals for the empirical statistics are calculated with the standard deviations and sample sizes from the actual data.

5.1 Factors affecting the overflow proportion

In this section, we discuss five factors that contribute to the challenges of fully reproducing the empirical overflow proportions.

Passive and intentional overflow. First, our model does not capture all overflow events happened at NUH. There are two kinds of overflow events in practice, which are triggered by different factors: *passive overflow* is triggered to avoid excessively long waiting times, while *intentional overflow* is triggered by other reasons such as medical needs. An example for an intentional overflow is that a General Medicine patient with a potential heart problem is overflowed to a Cardiology ward for telemetry care. See similar descriptions on intentional overflow in [8]. Our model captures passive overflow but not intentional overflow, whereas the empirical estimates include both overflow events. (Note that it is difficult to differentiate between passive overflow and intentional overflow from the data we currently have.) As a result, the empirical overflow proportions can be higher than the simulation estimates from our model.

Bed classes. Second, as elaborated in Section 4.1, NUH has various bed classes that can constrain bed assignments and affect the overflow proportion. Due to data unavailability we do not model bed classes in the present study. Instead, we use three overflow server pools to represent the wards having Class A/B1 beds. However, we do not expect such adjustment can fully capture the NUH operations.

Shared wards. Third, our model assumes the shared server pools 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 the shared wards, as indicated by the empirical study at NUH [7]. This complete flexibility in our model can reduce the occurrence of overflow events. For example, Neurology and General Medicine specialities share a ward (Ward 53 at NUH, corresponding to server pool 1 in the model). From the baseline simulation, Neurology patients constitute 29% of all primary admissions to the shared server pool. However, this proportion is only 18% in Period 1 from the empirical data. This low proportion suggests that sometimes a Neurology patient may not be able to be admitted to the shared ward even if a bed is available there. In this case, the Neurology patient could be overflowed to other wards in practice, but such an overflow does not occur in our model. As a result, the overflow proportions estimated from our model can be lower than the empirical estimates.

Other constraints in the bed allocation. Fourth, there are many other constraints that are considered in the real bed allocation process but not in our model, such as patient gender and infectious diseases. These constraints can also make the simulation estimates of overflow proportions lower than the empirical estimates. For example, in practice, when a female Surgery patient requests a bed but only a surgery bed for male is available, this patient might have to be overflowed to a Medicine bed. However, in our model, this patient will not be overflowed since the model does not differentiate between patient genders. Similarly, NUH generally does not put non-infectious patients in the same room with infectious patients.

Simplified overflow policy. Finally, the simple overflow policy used in the baseline model also prevents us from perfectly reproducing the overflow proportion. In practice, the overflow decision is extremely complicated, subjective, and often relies on individual patient's medical condition. In fact, there is no consistent guideline of how to overflow patients at NUH. The overflow policy we have used in the baseline is only a rough approximation of the practice we understand from NUH. We do not expect it to fully capture all overflow decisions at NUH.

In summary, the model we proposed in the main paper does not incorporate many constraints in the real practice at NUH. Given that our model is calibrated for the waiting time performance (see discussion in Section 5.1 of the main paper), not for the overflow proportions, we should not expect

our model to perfectly reproduce the overflow proportions. Still, the fact that overflow proportions from our simulation estimates turn out to be fairly close to the empirical estimates provides further confidence for our proposed model. In the next two sections, we conduct sensitivity analysis on the server pool setting and overflow polices to test the robustness of the insights generated in Section 6 of the main paper.

5.2 Sensitivity analysis on the server pool setting

Recall that each of the 15 server pools in our model corresponds to one or a set of similar wards at NUH, and we adjust the number of servers in the baseline model because the bed capacity had been increasing at NUH in Period 1. To understand how server pool setting affects the waiting time performance and overflow proportion, we test two alternative server pool settings in this section. We keep all other settings the same as in the baseline scenario.

In the first alternative setting, we no longer adjust the number of servers but directly use the number of beds at the *end* of Period 1 as the model input. Moreover, we do not assume any overflow wards. The three server pools with indexes 12-14, which were the three overflow wards in the baseline model (see Table 1 in the main paper), now operate as shared server pools. Based on their correspondence to the actual wards at NUH, they are designated to serve patients from Medicine and Cardiology (pool 12), Surgery, Orthopedic, and Oncology (pool 13), and General Medicine and Surgery (pool 14), respectively.

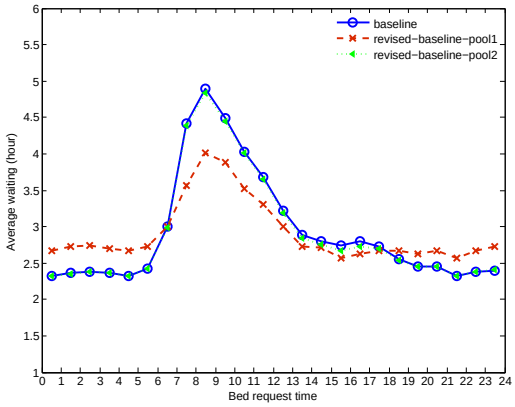
Since we do not adjust the number of servers in the first alternative setting, we need to increase the arrival rate to make the daily waiting times from simulation close to the empirical estimates; otherwise, the system utilization becomes too low because we use the bed capacity at the end of Period 1 (which is larger than the average capacity of Period 1). We increase the volume of arrival rate, uniformly across all specialties and admission sources, by 12%.

In the second alternative setting, we also directly use the number of beds at the end of Period 1 and increase the arrival volumes by 12%. However, we make two more adjustments. First, pools 12-14 are kept as overflow wards (as in the baseline setting). Second, we redistribute some servers from Orthopedic, Gastroenterology, and Renal pools to pools 12-14 to reflect the higher overflow rates in these wards from the empirical study; this is similar to the adjustment we made in the baseline setting. Comparing to the baseline model, in this setting, we increase the arrival volume to accommodate the changing capacity in Period 1 instead of reducing the number of servers in certain pools.

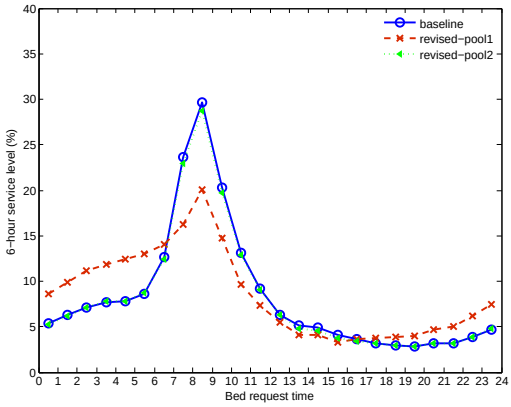
5.2.1 Waiting time and overflow proportions

We call the scenario using the first alternative server-pool setting (no overflow wards) the *revised-baseline-pool1* scenario. Similarly, we call the scenario using the second alternative server-pool setting the *revised-baseline-pool2* scenario. In the two revised baseline scenarios, only the server pool setting is changed; all other settings remains the same as in the baseline scenario.

Figure 28 compares the waiting time performance under the baseline scenario, the revised-baseline-pool1 scenario, and the revised-baseline-pool2 scenario. From the figure we can see that the waiting time performance under the revised-baseline-pool2 scenario is almost identical to that in the baseline scenario. However, under the revised-baseline-pool1 scenario, the waiting time curves appear to be flatter, i.e., the gap between the highest hourly average and daily average is smaller than that under the baseline scenario. This difference mainly comes from the effect of the overflow wards: these wards create more flexible capacity during the night, but accept very few patients in the morning under the given overflow policy (i.e., aggressive overflow during the night and conservative



(a) Average waiting time



(b) 6-hour service level

Figure 28: Hourly waiting time statistics under the baseline scenario and scenarios with different server pool settings. The arrival volume is increased by 12%; all other simulation settings are kept the same in each scenario. In the *revised-baseline-pool1* scenario, there is no overflow wards; while in the *revised-baseline-pool2* scenario, the overflow wards are kept but the total number of servers is obtained from the end of Period 1 without adjustment.

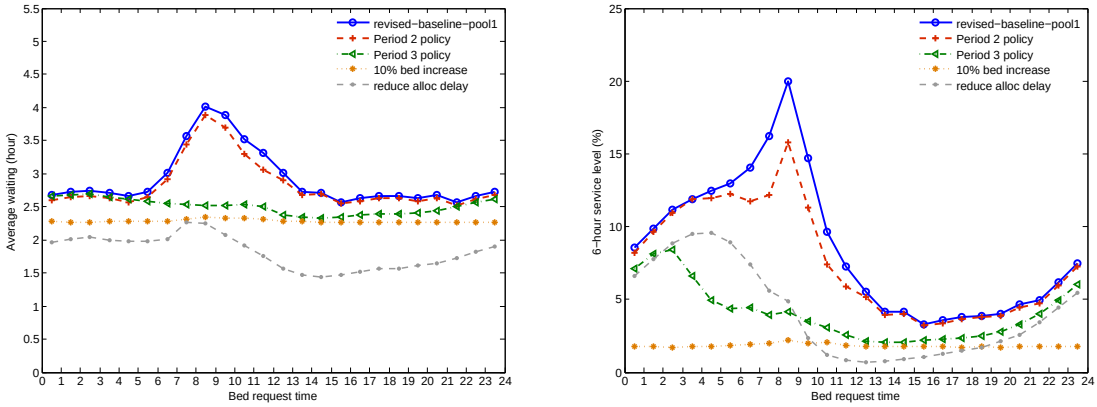
overflow during the day time). Thus, when overflow wards exist, the waiting time is shorter during the night but longer during the day, resulting in less flattened waiting time curves under the settings we have tested.

Moreover, the overflow proportion under the revised-baseline-pool1 scenario is only 4.81%, much lower than the empirical estimates (26.95%). In contrast, the overflow proportion under the revised-baseline-pool2 scenario is 19.07%, closer to that under the baseline scenario (21.70%) and the empirical estimates. This indicates that the adjustments we have done on the server pool setting, i.e., using the three overflow wards to mimic the admission control to Class A/B1 wards and redistributing some servers from Orthopedic, Gastroenterology, and Renal pools in the baseline scenario, are reasonable. Details of the adjustments are in Section 4.1.

5.2.2 Evaluation of five operational policies on the waiting time performance

In Section 6 of the main paper, we have gained various insights by evaluating five operational policies in the baseline server pool setting. To understand the robustness of these insights, we evaluate the five polices under each of the alternative server-pool settings. Figure 29 plots the hourly waiting time performance curves of four policies under the first alternative server-pool setting. In this figure, the curve corresponding to the Period 2 policy is simulated from a scenario that is identical to the revised-baseline-pool1 scenario except that the Period 1 discharge policy in the revised-baseline-pool1 scenario is replaced by the Period 2 discharge policy. Curves corresponding to other policies are obtained similarly. For comparison, we also plot the curve under the revised-baseline-pool1 scenario. Figure 30 plots the hourly waiting time performance curves for these four policies under the second alternative server-pool setting. Note that in each of the two alternative server-pool settings, the performance curves corresponding to the reduced LOS scenario are almost identical to those corresponding to the increased bed capacity scenario, and we do not plot them in the figures.

From these figures, we can see that the insights we gained in Section 6 of [6] still hold under the two alternative server-pool settings. First, Period 2 policy has limited impact on reducing



(a) Average waiting time

(b) 6-hour service level

Figure 29: Hourly waiting time statistics under the *revised-baseline-pool1* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the server pool settings are the same, i.e., we assume there is no overflow wards and arrival volume is increased by 12%. For Policy (ii) to (iv), the constant-mean allocation delay model is used.

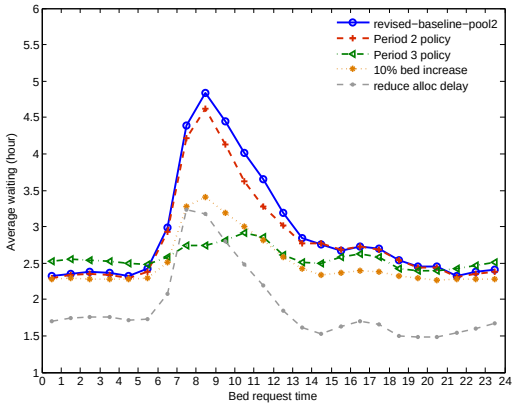
or flattening the waiting time statistics for ED-GW patients, whereas the hypothetical Period 3 policy can stabilize the hourly waiting time curves but has limited impact on the daily waiting time statistics. Second, increasing capacity, reducing LOS, or reducing mean allocation delays can reduce the daily waiting time statistics, but these policies alone do not necessarily stabilize the hourly waiting time performance.

5.3 Sensitivity analysis on the overflow policy

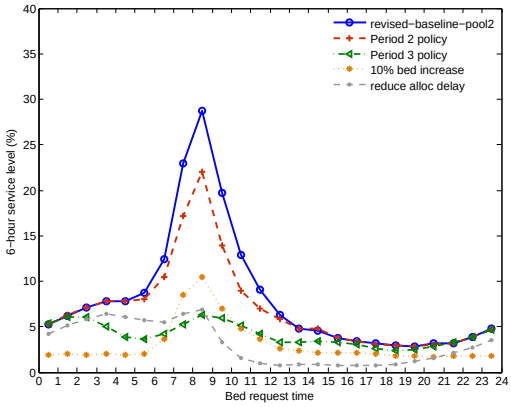
In the baseline scenario, we assume the overflow policy is fixed, and we focus on evaluating the early discharge policy and other policies under the given overflow policy. Clearly, the choice of overflow policy can greatly impact the waiting time performance; see Figure 11 and Figure 15 in the main paper for an example. As explained in Section 5.1, it is unlikely that one can model the actual overflow decisions at NUH exactly. Given this, we now test several alternative overflow policies for sensitivity analysis. Again, we keep all other settings unchanged from the baseline scenario.

We first test two extreme overflow policies. The first one is the full-sharing policy, i.e., the overflow trigger time T is always 0 and we overflow patients immediately if no primary bed is available. For the second policy, we assume there is no sharing ($T = \infty$) during the entire day, except that we do a full-sharing at the end of each day (midnight) to avoid unnecessary overnight waiting (i.e., to avoid the situation that a patient has to wait overnight when there is an overflow bed available). For convenience, we refer to the first policy the $T = 0$ *policy*, and the second one the $T = \infty$ *policy*. These two policies should produce the upper and lower bounds on the overflow proportion among a class of overflow policies that satisfy the following criterion: at least synchronize the entire system at the end of each day to avoid unnecessary overnight waiting. Note that the baseline overflow policy belongs to this class of overflow policies.

In addition, we test a third overflow policy, which is motivated by the $T = \infty$ policy and the baseline overflow policy. Under this policy, the overflow trigger time T is equal to 0.2 hour between 7pm and 7am the next day as in the baseline setting, and is equal to ∞ between 7am and 7pm as



(a) Average waiting time



(b) 6-hour service level

Figure 30: Hourly waiting time statistics under the *revised-baseline-pool2* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the arrival models are the same, i.e., we still keep the overflow wards but increase the arrival volume by 12%. For Policy (ii) to (iv), the constant-mean allocation delay model is used.

opposed to 5.0 hours in the baseline setting. We refer to this policy as the $T = \infty$-modified policy.

5.3.1 Waiting time and overflow proportions

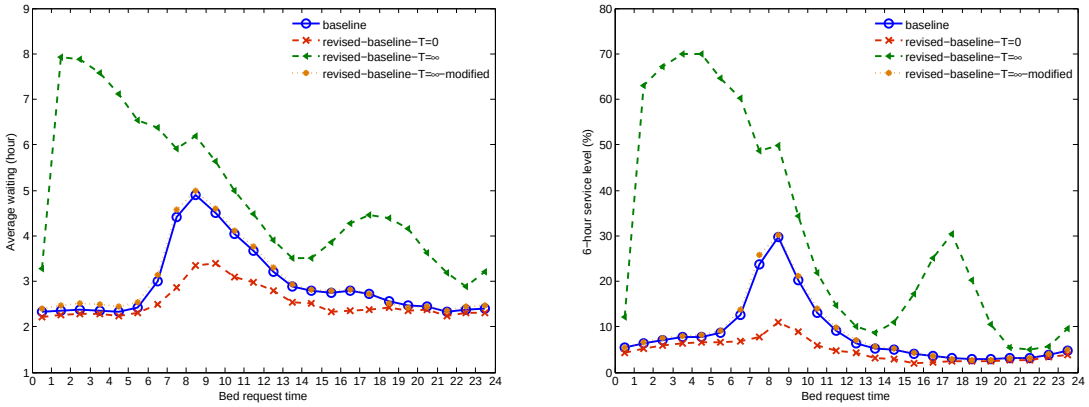
We call the scenario using the $T = 0$ policy the *revised-baseline-*$T = 0$ scenario. Similarly, we call the scenario using the $T = \infty$ policy and the $T = \infty$-modified policy the *revised-baseline-*$T = \infty$ scenario and the *revised-baseline-*$T = \infty$*-modified* scenario, respectively. In the three revised baseline scenarios, only the overflow policy is changed; all other settings remains the same as in the baseline scenario.

Figure 31 compares the waiting time performance under the baseline scenario and these three revised baseline scenarios. From the figure we can see that the waiting time curves under the revised-baseline-$T = 0$ scenario and the revised-baseline-$T = \infty$ scenario act as the lower- and upper-bound for the waiting time curves under the baseline scenario and the revised-baseline-$T = \infty$-modified scenario. This is expected because more resource sharing in the system leads to shorter waiting time. Moreover, note that the waiting time performance under the revised-baseline-$T = \infty$-modified scenario is almost identical to that under the baseline scenario.

It is not surprising that the overflow proportion under the revised-baseline-$T = 0$ scenario is the highest (22.75%), while the overflow proportion under the revised-baseline-$T = \infty$ scenario is the lowest (20.55%). The overflow proportion under revised-baseline-$T = \infty$-modified scenario is 21.70%, almost identical to the one under the baseline scenario. It appears that the overflow proportion is not very sensitive to the overflow policies under the settings we tested here.

5.3.2 Evaluation of five operational policies on the waiting time performance

To test the robustness of the insights generated in Section 6 of [6] against overflow policies, we evaluate the five operational policies under each of the three overflow policies introduced at the beginning of Section 5.3. In Figure 32, we fix the overflow policy to be the $T = 0$ policy and plot waiting time performance curves corresponding to the four operational polices specified in the figure



(a) Average waiting time

(b) 6-hour service level

Figure 31: Hourly waiting time statistics under the baseline scenario and scenarios with different overflow policies. All other simulation settings are kept the same in each scenario. In the *revised-baseline-*$T = 0$ scenario, a full-sharing $T = 0$ policy is used; in the *revised-baseline-*$T = \infty$ scenario, a $T = \infty$ policy (no sharing all day, do full-sharing at midnight) is used; and in the *revised-baseline-*$T = \infty$*-modified* scenario, a modified $T = \infty$ policy is used.

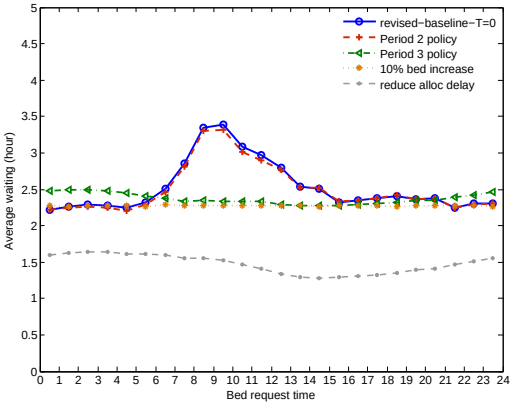
caption. Similar to what we did in Section 5.2, the curve corresponding to the Period 2 policy is obtained by simulating the scenario that is identical to the revised-Baseline-$T = 0$ scenario except that the Period 1 discharge policy is replaced by the Period 2 discharge policy. Other curves in the figure are produced similarly. In Figure 33, we fix the overflow policy to be the $T = \infty$ policy and plot waiting time performance curves. In Figure 34, we fix the overflow policy to be the $T = \infty$-modified policy and plot waiting time performance curves. Note that the performance curves under the reduced LOS scenarios are almost identical to those under the increased bed capacity scenarios, and we do not plot them in the figures.

From these figures, we can see that the insights we gained in Section 6 of [6] still hold under the tested alternative overflow policies. Since the performance curves under the $T = \infty$ policy do not have similar shapes as those under the baseline scenario, we interpret the findings in a slightly different way here. First, the Period 2 early discharge policy shows limited impact on reducing the waiting time statistics. In contrast, the hypothetical Period 3 policy shows a more significant impact on the waiting time performance. In particular, it can significantly reduce the waiting time for patients requesting beds in the morning (between 5am and 11am). Second, increasing capacity, reducing LOS, or reducing mean allocation delays can also help reduce the daily waiting time statistics.

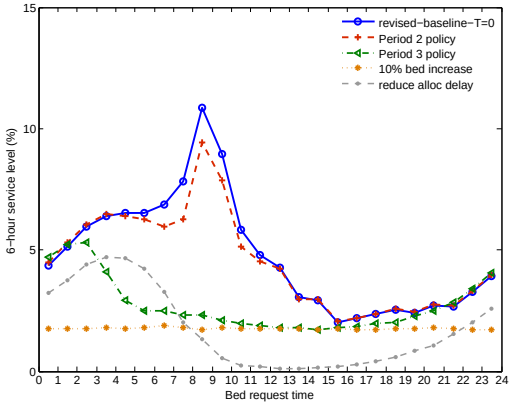
5.4 Early discharge policies have limited impact on the overflow proportions

Sections 6.1 and 6.2 of the main paper [6] demonstrate that early discharge policies have a limited impact on reducing the overflow proportion, even under the extreme midnight discharge policy. We provide an intuitive explanation for this observation here.

We consider patients requesting beds in the morning (7am to noon) since early discharge policies mainly affect these patients. In our simulation setting, the overflow trigger time T is long in the morning ($T = 5.0$ hours from 7am to 7pm). Thus, even in the baseline scenario, primary beds are likely to become available for morning arrivals before their waiting times exceed five hours (recall

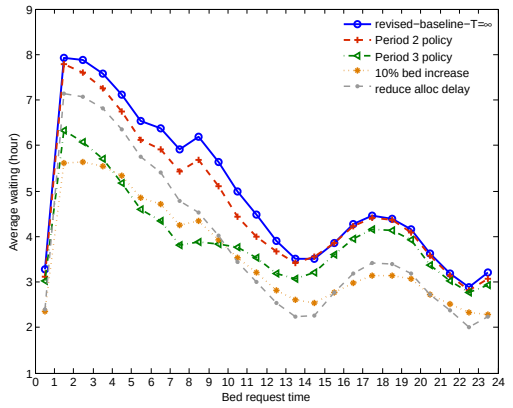


(a) Average waiting time

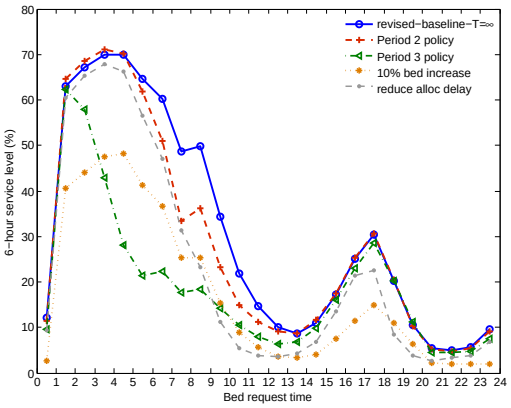


(b) 6-hour service level

Figure 32: Hourly waiting time statistics under the *revised-baseline-*$T = 0$ scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the overflow policies are the same, i.e., we use the full-sharing $T = 0$ policy. For Policy (ii) to (iv), the constant-mean allocation delay model is used.

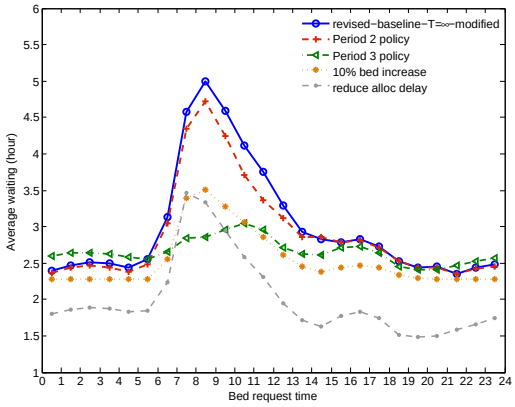


(a) Average waiting time

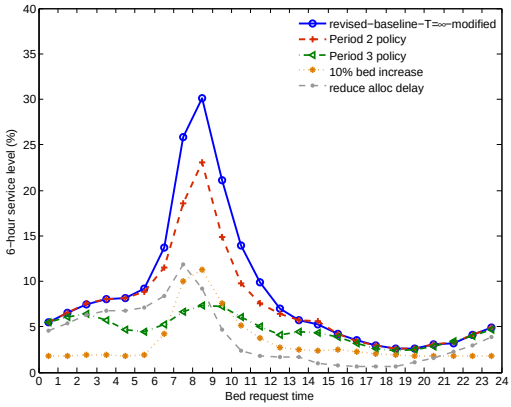


(b) 6-hour service level

Figure 33: Hourly waiting time statistics under the *revised-baseline-*$T = \infty$ scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the overflow policies are the same, i.e., we use the $T = \infty$ policy. For Policy (ii) to (iv), the constant-mean allocation delay model is used.



(a) Average waiting time



(b) 6-hour service level

Figure 34: Hourly waiting time statistics under the *revised-baseline-$T = \infty$-modified* scenario and scenarios with (i) Period 2 policy, (ii) Period 3 policy, (iii) 10% bed capacity increase, and (iv) reduce mean allocation delays. In all scenarios, the overflow policies are the same, i.e., we use the modified $T = \infty$ policy. For Policy (ii) to (iv), the constant-mean allocation delay model is used.

that most discharges start to occur from noon under the baseline Period 1 discharge distribution). In other words, under the given overflow policy, there are already very few morning arrivals overflowed in the baseline scenario. This is also indicated by the observation that the overflow proportion under the baseline overflow policy is very close to that under the $T = \infty$-modified policy, which do not allow any overflow between 7am and 7pm (see Section 5.3.1). Therefore, although more beds become available in the morning after the early discharge, the overflow proportion will not be affected much.

References

[1] M. Armony, S. Israelit, A. Mandelbaum, Y. Marmor, Y. Tseytlin, and G. Yom-Tov, "On patient flow in hospitals: A data-based queueing-science perspective," 2014, working paper. [Online]. Available: http://iew3.technion.ac.il/serveng/References/Armony_et_al_Patient_flow.pdf

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

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

[4] 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.

[5] P. Shi, "Stochastic Modeling and Decision Making in Two Healthcare Applications: Inpatient Flow Management and Influenza Pandemics," Ph.D. dissertation, Georgia Institute of Technology, 2013. [Online]. Available: https://smartech.gatech.edu/handle/1853/50367

[6] P. Shi, M. Chou, J. G. Dai, D. Ding, and J. Sim, "Models and Insights for Hospital Inpatient Operations: Time-Dependent ED Boarding Time," *Management Science*, 2014, forthcoming.

[7] P. Shi, J. G. Dai, D. Ding, J. Ang, M. Chou, J. Xin, and J. Sim, "Patient Flow from Emergency Department to Inpatient Wards: Empirical Observations from a Singaporean Hospital," 2014. [Online]. Available: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2517050

[8] 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.