The Joint Commission Journal on Quality and Patient Safety 2021; 47:354–363

Use of a Novel Patient-Flow Model to Optimize Hospital Bed Capacity for Medical Patients

Yue Hu; Jing Dong, PhD; Ohad Perry, PhD; Rachel M. Cyrus, MD; Stephanie Gravenor, MBA; Michael J. Schmidt, MD

Background: There is no known method for determining the minimum number of beds in hospital inpatient units (IPs) to achieve patient waiting-time targets. This study aims to determine the relationship between patient waiting time–related performance measures and bed utilization, so as to optimize IP capacity decisions.

Methods: The researchers simulated a novel queueing model specifically developed for the IPs. The model takes into account salient features of patient-flow dynamics and was validated against hospital census data. The team used the model to evaluate inpatient capacity decisions against multiple waiting time outcomes: (1) daily average, peak-hour average, and daily maximum waiting times; and (2) proportion of patients waiting strictly more than 0, 1, and 2 hours. The results were published in a simple Microsoft Excel toolbox to allow administrators to conduct sensitivity analysis.

Results: To achieve the hospital's goal of rooming patients within 30 to 60 minutes of IP bed requests, the model predicted that the optimal daily average occupancy levels should be 89%–92% (182–188 beds) in the Medicine cohort, 74%–79% (41–43 beds) in the Cardiology cohort, and 72%–78% (23–25 beds) in the Observation cohort. Larger IP cohorts can achieve the same queueing-related performance measure as smaller ones, while tolerating a higher occupancy level. Moreover, patient waiting time increases rapidly as the occupancy level approaches 100%.

Conclusion: No universal optimal IP occupancy level exists. Capacity decisions should therefore be made on a cohort-by-cohort basis, incorporating the comprehensive patient-flow characteristics of each cohort. To this end, patient-flow queueing models tailored to the IPs are needed.

Hospitals are constantly examining how to provide the right care, in the right place, at the right time. To achieve satisfactory outcomes, it is imperative that patients be admitted in a timely manner to the unit that is most appropriate for them. This, in turn, requires the hospital to have sufficient capacity in each inpatient and observation unit/cohort (IP) to satisfy the demand. Indeed, chronic shortage of IP capacity results in overcrowding and boarding of patients in the emergency department (ED) and other areas, which puts patients at risk for suboptimal care and potential harm.[1–4] Inadequate capacity can also increase the burden on hospital staff and accelerate clinician burnout.[5] Thus, hospital administrators are faced with the difficult task of balancing the trade-offs between the high costs associated with increasing the bed capacity and the need to ensure timely care. The fundamental challenge lies in the lack of a comprehensive framework for understanding the impact of resource allocation decisions on patient-flow outcomes.

In hospital capacity planning, particular attention must be given to the resource allocation in the IPs, which, due to their central location (that is, as hubs) in the hospital network, have large impacts on upstream units, such as the ED, ICUs, postanesthesia care unit (PACU), operating rooms (ORs), and external direct admission sources.[6–9] Congestion in the IPs tends to propagate upstream because patients who need to be admitted to the IPs cannot be transferred until IP beds become available. Thus, insufficient capacity in the IPs can lead to ED and PACU boarding, which in turn may cause further undesirable outcomes, including patients leaving the ED without being seen, ambulance diversion, and OR holding.[7,10]

The ideal occupancy in the IPs is often stated to be such that daily average occupancy is about 85%.[11–16] One of the earliest expressions of this occupancy can be traced to the discrete-event simulation implemented in Excel by Bagust et al.[11] The authors report that "risks are discernible when average bed occupancy rates exceed about 85%, and an acute hospital can expect regular bed shortages and periodic bed crises if average bed occupancy rises to 90% or more."[11(p. 155)] According to Bain et al., the emergency medicine in Australia directly advocated 85% occupancy when addressing the overcrowding problem: "Queuing theory developed by Erlang nearly 100 years ago tells us that systems are most efficient when they operate at 85% capacity. This applies to queues at the local bank waiting for the teller or at ticket booths at the MCG [Melbourne Cricket Ground]."[12(p. 42)] Others have called for similar 85% occupancy.[13,14]

1553-7250/$-see front matter

https://doi.org/10.1016/j.jcjq.2021.02.008

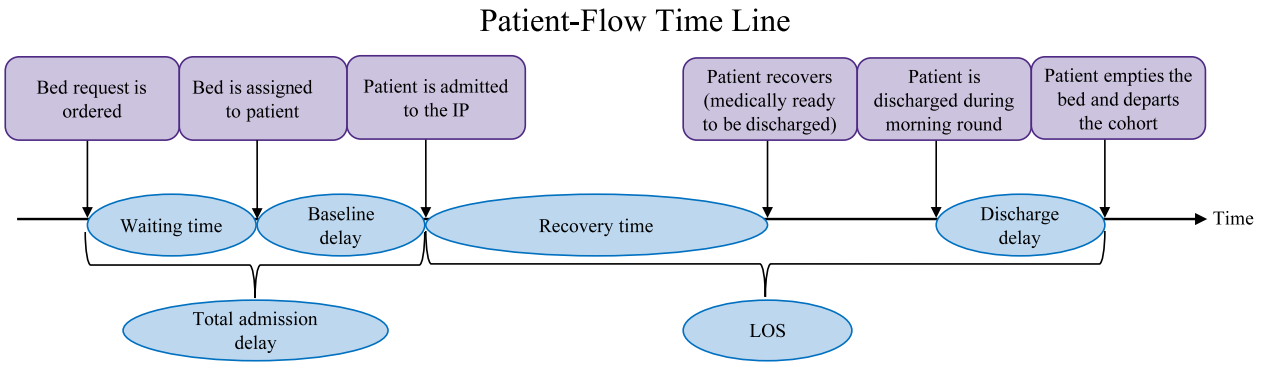

Figure 1: Shown here is a typical patient-flow timeline in an inpatient unit (IP) cohort. LOS, length of stay.

We are unaware of a general result (that is, rule or theorem) in queueing theory claiming that systems are most efficient when operating at 85% utilization; indeed, this level of utilization is often too low for large systems with many servers and too high for small systems with a small number of servers. The lack of a universal optimality rule can partially explain why different studies found different optimal occupancy rates, ranging from 70% to 90%.[17–19] Furthermore, as was mathematically demonstrated in Dong and Perry, existing queueing models (such as Erlang's), do not accurately capture the complex operational characteristics of IP cohorts and the patient-flow dynamics therein, necessitating the development of specialized models.[20]

We sought to address this gap by implementing a novel queueing model specifically designed for hospital IPs in Dong and Perry[20] to analyze the relationship between the occupancy levels of a unit (or cohort of beds) and timeliness of obtaining a bed in the cohort. We quantified the minimum number of beds needed to achieve patient waiting time–related performance measures for each IP. In particular for the Medicine, Cardiology, and Observation cohorts of a large urban academic hospital (which houses 400 private IP rooms and 92 private ICU rooms), hospital management set the patient waiting-time performance goal at less than 60 minutes, defined as the time difference between bed request and bed assignment. To this end, we simulated the queueing model for each cohort, validated the model's accuracy by comparing its prediction to patient-flow data, and determined the optimal number of beds needed. To facilitate use of the model without expertise in queueing theory and simulation techniques, we published the results in a simple Microsoft Excel toolbox that would allow administrators to conduct sensitivity analysis and evaluate various bed allocation policies. The toolbox could be calibrated to other IP cohorts and hospitals as well, using the same input data as described in the "Model Input" section below.

METHODS

Queueing theory is concerned with the modeling and analyses of randomly evolving systems that process work, such as communication, manufacturing, inventory, and service systems, among others.[21–28] In our IP setting, the "work" is the clinical care of patients, and the "servers" are the beds in the cohort under consideration. Patients are considered to be "queueing" (waiting for service) if their bed requests have been ordered but no bed has yet been made available. Those patients are waiting for their requested beds in other units of the hospital, with most of them boarding in the ED. A patient's waiting time is then the period between the time at which the bed request is made and the assignment time of the patient to an available IP bed in the desired service.

Model Description

We simulated the queueing model in Dong and Perry[20] using Monte Carlo methods,[29,30] which take into account the salient patient-flow characteristics of hospital IP cohorts. The dynamics in the queueing model are as follows (see Figure 1 for a graphical representation): Bed requests to the IPs are entered with a rate that is equal to the average number of bed requests per hour. If a bed request is made while there are available beds in the IP, the patient for whom the bed is requested is assumed to be assigned that bed immediately. If no bed is available when a request is made, the corresponding patient waits until a bed becomes available. If multiple patients are waiting for a bed in the same IP, the bed is assigned in the order that it was requested (first in, first out). After the patients have recovered, they remain in their beds until the next morning physician round in which they are identified as recovered and given a disposition order to be discharged. However, the actual departure does not take place immediately after the morning round but after additional administrative steps and delays. Those latter delays may be due to paperwork, transportation arrangement, patient education by the care team, and so on and may take several hours for some patients. We term the time difference between the discharge order and the actual departure of a patient as the "discharge delay." The recovery time and the discharge delay together add up to the patient's length of stay (LOS) in the IP. The departure process just described, in which hospitalized patients depart only after discharge decisions are made in the morning round, with subsequent

Table 1. Inpatient Unit (IP) Delay Measures

Measure	Start	End	Cause	Performance goal?
Bed assignment delay (waiting time)	Bed request order time	Bed assignment time	IP bed is located and assigned to another patient	Yes, 30–60 minutes on average
Post-assignment delay (baseline delay)	Bed assignment time	IP admission time	Physician report/handoff, bed cleaning completion, transportation, etc.	No, 29 minutes on average
(Total) admission delay	Bed request order time	IP admission time	Sum of bed assignment delay and post-assignment delay	–
Discharge delay	Physician discharge order time	Patient departure time	Paperwork, transportation arrangement, patient education, etc.	No

discharge delays, is unique to our model. Further, it is closer to reality than any other discharge process in the queueing literature. It is therefore significant that this process has substantial impact on the patient-flow dynamics, implying that standard queueing models are not appropriate for IP modeling. Our novel queueing model allows for generalizability to other hospitals and cohorts that satisfy the key model assumptions: (1) (mostly) unscheduled patient demand, (2) random bed request times and LOS, (3) physician morning round, and (4) discharge delay.

We make three remarks on the validity of the aforementioned model assumptions to capture reality. First, patients can experience further delay after being assigned an IP bed. We refer to this post-assignment delay (in comparison to the bed assignment delay) as the baseline delay, which can include the time managers need to communicate with the environmental services team, the travel time taken by the servers to move to the assigned bed location, the bed cleaning time, and the time required to finish duties associated with any prior assignment. At our hospital, the average baseline delay is approximately 29 minutes. The bed assignment delay and baseline delay add up to the total admission delay of a patient (see Table 1 for a summary of IP delay measures). That said, it is important that our goal is to control the bed assignment delay, which is the time between the bed request and bed assignment, to be between 30 and 60 minutes, independent of the baseline delay. Second, it is possible in practice (and not explicitly captured by the model) that patients who are more acutely ill would be preferentially admitted out of order. Nevertheless, assigning different priorities and changing the sequence of the queueing patients does not affect the average waiting time measure. Thus, the model is effective in making capacity decisions regardless of the prioritization of patients. Third, when a patient waits an excessively long time before a primary bed becomes available, hospital managers may choose to assign them to a nonprimary bed, which is often referred to as off-service placement. Although the practice of off-service placement exists, we do not allow it in the model when determining the optimal bed capacity, as the goal is for patients to be assigned to the medical service unit most appropriate to their clinical needs. Indeed, off-service placement has been shown to be associated with worse medical outcomes and longer LOS, which are undesirable from both the clinical and the efficiency perspective.[31] In our analysis, the bed capacity is set such that off-service placement does not occur during the normal state of operation. Therefore, the waiting times presented in this report represent those for ideal bed assignment instead of those with off-service placement.

Model Input

At our hospital, the model was found to appropriately describe the bed request and discharge patterns in the Medicine, Cardiology, and Observation cohorts, because most admissions to these cohorts are unscheduled (mainly from other hospital units or the ED), and each of these cohorts had the discharge pattern described above (morning round and discharge delays). Because the hospital geographically localizes clinical cohorts to concentrate specialized care teams for the patients, these cohorts have a fixed number of beds with minimal overflow to other cohorts.

For each cohort, we fit the input parameters for the model from the hospital's data. These include (1) the average number of bed requests per hour (arrival rate), (2) the average LOS of patients in the IP, (3) the average number of beds that become available (departure rate) per hour on a typical day, and (4) the typical time in the morning at which the round ends. We make two comments on the model input. First, departure rate in the model is reflected by the average number of beds that become available per hour on a typical day (input 3), which can be estimated from the patient departure time stamps in the data. For the discharge delay, if 10% of patients depart the unit 2 hours after the morning round on average, then the discharge delay is assumed to be 2 hours with probability 0.1. Second, we use the typical end time of the morning round (input 4) to approximate the average discharge order time. Alternatively, hospital management can calculate the mean of discharge order times across all patients from data and use this value for input 4.

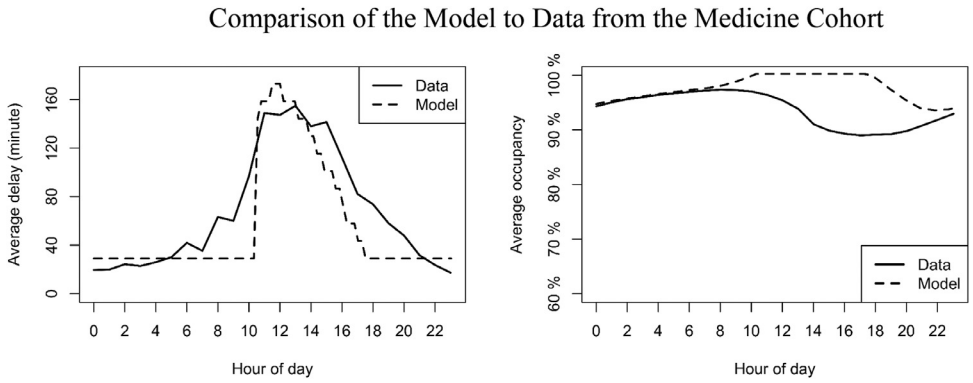

Figure 2: The two graphs compare the model to data from the Medicine cohort. Left: Average admission delay for a bed in the Medicine cohort. Right: Hourly occupancy rate in the Medicine cohort. The model is validated retrospectively for satisfied inpatient bed demand (that is, excluding unmet demand).

Model Output

The model can be used to evaluate patient flow–related performance metrics at any bed capacity level. Focusing on controlling patients' average waiting time for an IP bed, we computed five performance measures for any given number of beds: (1) hourly occupancy level in the IP cohort, (2) average waiting time for bed requests in the IP on a typical day, (3) peak-hour waiting time (that is, waiting time for bed requests made during the busiest hour of the day), (4) daily maximum waiting time (that is, the maximum waiting time for bed requests on a typical day), and (5) the proportion of patients who have to wait due to unavailability of beds upon the request, as well as the proportion of patients who wait strictly more than 1 hour or more than 2 hours for an IP bed. Given desired waiting-time performance constraints, such as keeping the average waiting time within the range of 30 to 60 minutes, the model can then be used to determine the number of beds (and the corresponding occupancy level) needed to achieve the desired service constraint.

Model Validation

We validate the accuracy of the queueing model proposed by Dong and Perry[20] for our hospital by calibrating the model input from data and comparing the model output to the observed dynamics. For the Medicine cohort, with its 172 beds, this comparison is depicted in Figure 2. We observe that the hourly total admission delay predicted by the model matches the data closely. In particular, bed requests are sporadically made early in the morning (before 10:00 A.M.), when the queue for the inpatient beds is negligible. Thus, admission delays during those hours are taken to be the average baseline delay (29 minutes) in the model. The majority of bed requests are made by the upper stream units between 10:00 A.M. and 6:00 P.M. Due to the large volume of bed requests and the discharge delay in the patient departure process, patients start to queue for the inpatient beds and experience significant waiting times, beyond the baseline delay. After 6:00 P.M., all the bed requests are eventually accommodated, as all the discharged patients have left the unit, so that the model-predicted admission delay is again equal to the average baseline delay.

We follow the same rationale as in regression analysis to calculate a coefficient of determination for our model. In particular, the total sum of squares of the hourly average admission delay (SS_{tot}, proportional to the variance of the observed data) is 53,824.25, and the sum of squares of the residuals of the model (SS_{res}) is 11,596.85. Thus the coefficient of determination (defined as $1 - SS_{res}/SS_{tot}$) is equal to 0.785. This means that 78.5% of the hourly variation in the admission delay throughout the day can be explained by the model.

However, there is discrepancy between the model-predicted and observed bed occupancy process. As shown in the left graph in Figure 2, the observed average occupancy never reaches 100%, suggesting that there is always spare capacity on average. In contrast, the right graph indicates long admission delay during most of the day, particularly in the time window when the observed occupancy level is the lowest. Given the long admission delay during the day, the underrepresentation of midday occupancy in the data can be attributed to census inaccuracy during turnover of beds or temporary closure of beds for maintenance, which may falsely suggest less than 100% occupancy. Note that the census issue related to the occupancy data is restricted to the operations of our hospital and may not be generalizable to other sites. On the other hand, our model predicts 100% occupancy during the hours when patients experience excessive delays.

Model Implementation

The simulation of the queueing model was coded into an Excel toolbox which can be easily used by hospital management. In addition, to prepare the hospital for projected future demand, we incorporated the forecasted compound annual growth rate for demand in the model so that it could be used to determine future capacity needs

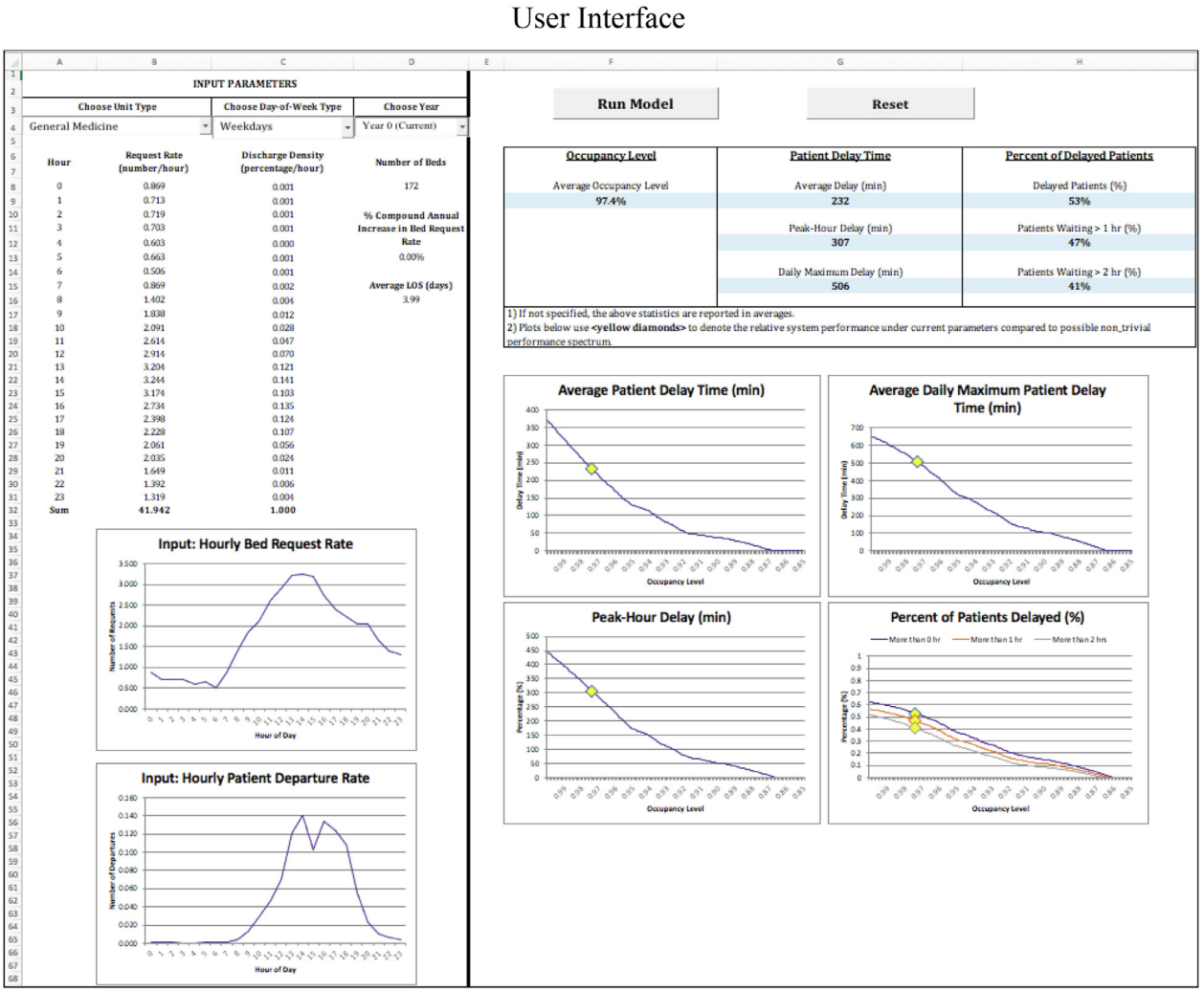

Figure 3: Shown here is the user interface of the Excel toolbox. (See Appendix 1, available in online article, for guidance to download and use the Excel toolbox.)

as market conditions change. The user interface of the toolbox is shown in Figure 3. This toolbox can be easily employed by any hospital to analyze units/cohorts that adhere to the modeling assumptions, by simply feeding it with the required data as input. The Excel toolbox is openly accessible as Appendix 1 (available in online article), or through a GitHub repository.[32]

RESULTS

Relationship of Patient Waiting Time and IP Occupancy Level

Simulation results of our model provide the average waiting time, peak-hour waiting time, and daily maximum waiting time for different bed capacity levels (Figure 4). For example, the Medicine cohort achieves an average occupancy level of 92.0% with 182 beds, for which the resulting average delay is 54.3 minutes, the peak-hour delay is 77.6 minutes, and the daily maximum delay is 158.2 minutes. We also obtain the proportion of patients who would wait for more than 1 and more than 2 hours in correspondence to difference bed capacities (Figure 5). At 92.0% average occupancy level, it is estimated that 16.3% of patients would wait for more than 1 hour, and 12.6% of patients would wait for more than 2 hours for a bed in the Medicine cohort. Figure 5 also demonstrates that the marginal degradation in the waiting-time performance measures becomes more rapid as the number of beds decreases and the average occupancy level increases.

Optimal Bed Capacity Decisions

With the goal of an average waiting time between 30 and 60 minutes, the simulated model determined target average occupancy levels of 89%–92% for the Medicine cohort, 74%–79% for the Cardiology cohort, and 72%–78% for the Observation cohort (Table 2). Table 2 also shows the current number of beds in each cohort, the current average occupancy and waiting times, and the number of beds that need to be added to each cohort to achieve the target waiting times. The desired occupancy levels can be determined

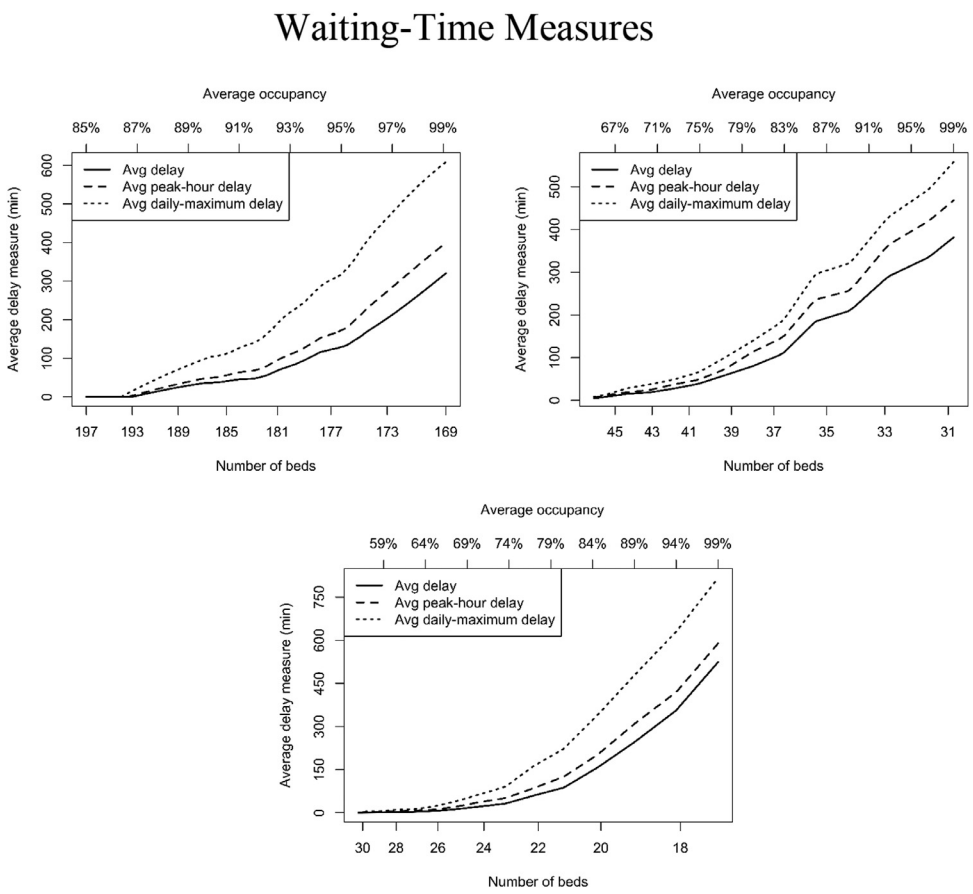

Figure 4: The graphs show waiting-time measures for the Medicine (top left), Cardiology (top right), and Observation (bottom) cohorts. The performance spectrum is derived for projected inpatient bed demand informed by hospital management.

Table 2. Bed Capacity Decision for the Medicine, Cardiology, and Observation Cohorts

Cohort	Current beds	Current avg. occupancy	Current avg. waiting time	Target avg. occupancy*	Number of beds to add
Medicine	172	97%	232 (min)	89%–92%	10–16
Cardiology	32	97%	333 (min)	74%–79%	9–11
Observation	20	85%	166 (min)	72%–78%	3–5

* The target occupancy is set to achieve 30–60 minutes average wait time.

similarly for other target performance measures. For example, if hospital management aims to control the peak-hour delay to be less than 30 minutes, the target average occupancy level is 88% for the Medicine cohort, 71% for the Cardiology cohort, and 69% for the Observation cohort.

Identification of Critical Capacity Levels or Capacity Tipping Points

Patient waiting time increases rapidly as the occupancy level approaches 100% in each IP cohort. We identify *tipping point* as the critical occupancy level above which service quality degrades sharply. Formally, we define the marginal increase (alternatively, the growth rate) in the average waiting time after removing 1 bed when the total number of beds is N as

$$S(N) = (\text{average waiting time with } N-1 \text{ beds}) - (\text{average waiting time with } N \text{ beds})$$

Tipping point is the occupancy level at which the marginal increase in the average waiting time first exceeds 30 minutes. Consider the Cardiology cohort in particular (Figure 4). As the number of beds decreases from 36 (85% average occupancy) to 35 (87% average occupancy), the average waiting time increases by 20 minutes. As the number of beds further decreases from 35 (87% average occupancy) to 34 (90% average occupancy), the average waiting time increases by 49 minutes. Thus, we identify the average occupancy level of 87% with 35 beds as the tipping point for

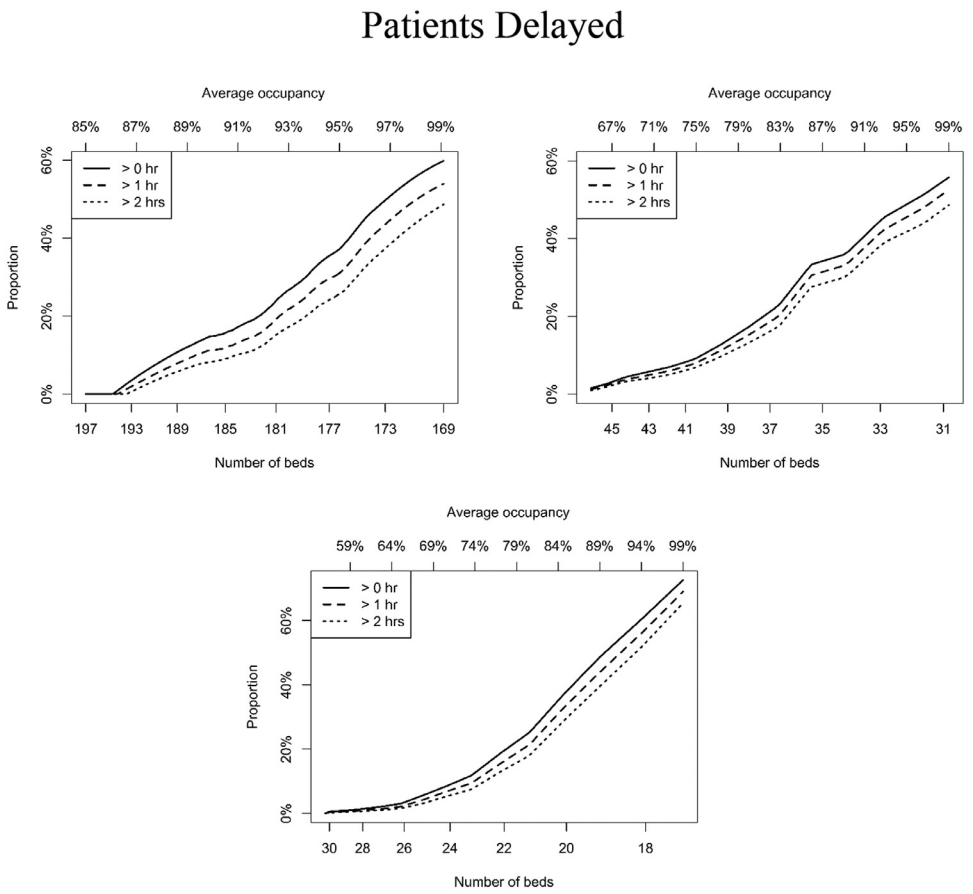

Figure 5: The graphs show proportions of patients delayed for the Medicine (top left), Cardiology (top right), and Observation (bottom) cohorts. The performance spectrum is derived for projected inpatient bed demand informed by hospital management.

the Cardiology cohort. Similarly, the tipping point is 99% with 169 beds for the Medicine cohort, and 85% with 20 beds for the Observation cohort.

Capacity Expansion Outcome

Based on the model recommendations, we elected to expand the Medicine bed capacity by a total of 20 beds by September 2018 (+8 beds on August 20, 2018 and +12 beds on September 10, 2018), and in tangent to that also expanded provider and nurse coverage for this hospital service line. A direct outcome was the reduction in ED ambulance diversion, which was used in times of severe overcrowding to reduce demand and mitigate congestion (Figure 6). Prior to expanding beds for the Medicine unit, diversion hours for our hospital were as high as > 50% of the hours in a month. As a consequence of the reduced waiting times and occupancy levels that the novel patient-flow model correctly predicted, our hospital was able to sustain the reduction in ambulance diversion to < 20% of the hours in October and November 2018.

DISCUSSION

Understanding the relationship between hospital occupancy level and queueing-related performance measures is a complex problem confronting hospital and health system management. Despite continuous efforts in gauging the ideal occupancy that meets a desired service quality, overcrowding remains common across hospitals and is expected to worsen.[13,33]

To address the capacity-planning problem in hospitals, we used a high-fidelity queueing model specifically designed for IPs that employ morning rounds. Our model successfully retrieves the one-to-one relationship (that is, trade-off curves) between the waiting time–related performance measures and the occupancy level. Understanding this interplay can inform hospital administrators on the correct target occupancy level to achieve the desired quality-of-care target. The insights and results gained are summarized below.

1. There is no universal occupancy level that can be considered ideal across IP cohorts, for two reasons. First, different patient waiting time performance targets can lead to different desired occupancy levels. For example, the occupancy required to keep the average waiting time below a certain threshold may be different from the occupancy required to keep the probability of patients waiting below a desirable value. Second, and more important, the occupancy required to achieve a fixed performance mea-

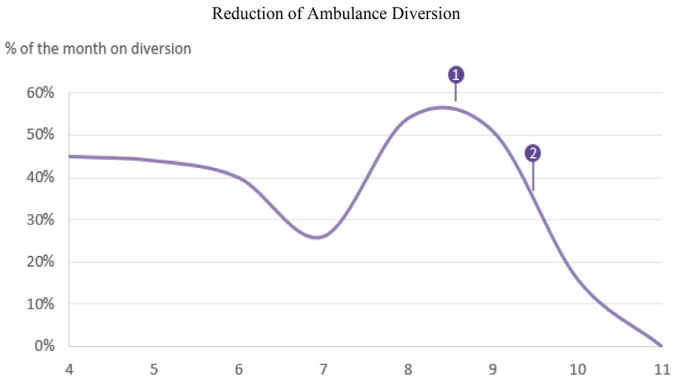

Figure 6: The graph shows the reduction of ambulance diversion after capacity expansion. Eight Medicine beds had been added by August 20, 2018 (1), and an additional 12 beds had been added by September 10, 2018 (2).

sure depends on other characteristics of the cohort, such as the average LOS and the bed request rate (patient demand). Two different cohorts can have two very different ideal occupancies, even if they both target the exact same performance metric. Indeed, as we demonstrate for our hospital, to achieve an average waiting time of about 30 to 60 minutes, the Observation cohort should have its occupancy at about 70%, whereas the Medicine cohort should have an occupancy of approximately 90% (Figure 5). Thus, no rule of thumb, such as the 85% occupancy rule, can be used for all cohorts. An optimal occupancy must be determined on a case-by-case basis, depending on the specific characterization of the IP and its patient flow dynamics and on the performance metric goals management seeks to achieve. For example, with everything else held constant, it follows from classical queueing theory that cohorts with smaller variability in LOS may achieve the same waiting time performance with fewer beds.

2. It can be mathematically verified that larger cohorts with higher patient volume can run at a higher occupancy level while achieving the same queueing-related performance measures as smaller ones with fewer patients. Indeed, as can be seen in Figure 5, the Medicine cohort (188 beds) can achieve 30 minutes average waiting time with 89% occupancy, whereas the same performance requires a 74% occupancy in Cardiology (43 beds), and a 72% occupancy in Observation (25 beds). This empirical verification is a demonstration of the economies-of-scales principle for queueing systems, which stipulates that service efficiencies increase directly as the system scale increases. In particular, the randomness in patient-flow dynamics plays a smaller role in larger systems than in smaller ones. The insight for hospital administrators is twofold. First, at least as far as patient flow is concerned, there are operational advantages (for example, reducing waiting time for bed requests) to pooling cohorts (beds) together when this is clinically and operationally feasible. Second, expanding capacity for a small cohort may lead to larger marginal gain in performance stabilization than expanding capacity for a large cohort. (In reality, prioritization of increasing capacity may also be given to cohorts that are more frequently selected as the off-service placement destinations. This latter aspect is not considered in the scope of this paper.)
3. As shown by Figures 4 and 5, patient waiting time and the proportion of patients waiting for a bed increase rapidly as the occupancy level approaches 100%. This suggests that hospitals operating at a high occupancy level (close to full bed utilization) will experience drastic deterioration in service timeliness with slight increases in demand. Recall that the trade-off curves (Figure 5) suggest that the critical occupancy levels are approximately 99% for Medicine, 87% for Cardiology, and 85% for Observation. These are occupancy levels at which the change in the growth rate of the average waiting time exceeds 100% if one additional bed is removed. (The fact that the tipping point is larger for larger cohorts is another indication of the economies of scales discussed above.) Comparing these critical occupancy levels to the target occupancy levels, we observe that the increase in bed capacity recommended by the model will lead to improved and robust system performance, in that variability plays less of a role in pushing the IPs into severe congestion.

Strengths of the Model

In practice, the queueing model is easy to implement and relatively accurate. Hospital management used the model to make a data-driven decision to increase Medicine, Cardiology and Observation beds at the hospital to improve bed availability and reduce patient waiting times. Our Excel program for the model is generalizable to other hospitals and cohorts within similar patient-flow characteristics described in the Methods section. Furthermore, the model can be used to conduct sensitivity analysis that is not in the scope of this paper. For example, given that discharge delay negatively affects the patient-flow process because a discharged patient can block an incoming patient by occupying the bed after being discharged, the model can be used

to quantify the impact of shortening discharge delays and discharging patients earlier in the day.

Limitations of the Model

General limitations of this model include that it does not directly incorporate seasonal variations (note, however, that the model does allow hourly and daily variations). For hospitals where seasonality is salient, this issue can be circumvented by analyzing each season of the year separately. In addition, the queueing model assumes that the decision to discharge a patient is made in the morning round. In reality, some discharge decisions are made later in the day, prompting continuous, rather than one-time, discharge decisions. When the number of discharge decisions that are taken outside the morning round is relatively small, the queueing model is accurate. If discharge decisions outside the morning round are common, the model provides a worst-case scenario on the achievable performance metrics.

In addition, although our toolbox assumes exponential distribution for LOS (as in traditional queueing models), existing empirical evidence suggests that LOS in IPs may be more accurately described by lognormal distributions.[28] The simulation can be refined to include other hospital-specific operational features (including but not restricted to the continuous physician rounding and LOS distribution) not captured by the current model.

As mentioned in the Methods section, assigning different priorities and changing the admission order of the waiting patients does not affect the *average* waiting time measure. Thus, the model is effective in making capacity decisions on a macroscopic level, regardless of the prioritization of patients. However, the limitation is that the model does not take into account more granular patient-level performance targets; for example, waiting time for acute patients when acute patients can be admitted out of order to an IP bed.

Last, the queueing model we have implemented is limited to settings where all beds are in private rooms. Bed assignment decisions can be much more complex in settings where the infection precaution is high, and the number of shared rooms is large. To the best of our knowledge, how to determine the number of beds (single vs. shared) and make the corresponding bed assignment decisions effectively is based on experience. We identify this problem as an interesting direction for future research.

CONCLUSION

We described the use of a high-fidelity queueing model that accurately captures the trade-offs between patient waiting time–related performance measures and IP bed utilization. Given desired operational performance constraints (such as average waiting times and proportion of patients waiting for a bed), that queueing model can be used by hospital management to derive the optimal number of beds—namely, the minimum number that satisfies the desired service-level constraints. The model's computational algorithms were coded into an easy-to-use Excel toolbox that can be used by any hospital with similar IP operational characteristics and patient-flow dynamics—namely, (1) (mostly) unscheduled patient demand, (2) random bed request times and LOS, (3) physician inspection round, and (4) discharge delay. The toolbox was employed by a large urban academic hospital to recommend the number of additional beds that are needed to constrain the average patient wait times to 30–60 minutes in the Medicine, Observation, and Cardiology cohorts.

Funding. Jing Dong is partially supported by National Science Foundation (NSF), CMMI 1762544. Ohad Perry is partially supported by NSF, CMMI 1763100.
Conflicts of Interest. All authors report no conflicts of interest.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jcjq.2021.02.008.

Yue Hu is PhD Candidate, Decision, Risk, and Operations (DRO), Columbia Business School, New York City. **Jing Dong, PhD**, is Associate Professor of Business, DRO, Columbia Business School. **Ohad Perry, PhD**, is Associate Professor, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois. **Rachel M. Cyrus, MD**, is Associate Professor, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago. **Stephanie Gravenor, MBA**, formerly Project Manager, Department of Emergency Medicine, Northwestern University Feinberg School of Medicine, is Co-Founder and CEO, Medecipher, Inc., Denver. **Michael J. Schmidt, MD**, is Associate Professor, Department of Emergency Medicine, Northwestern University Feinberg School of Medicine. Please address correspondence to Yue Hu, yhu22@gsb.columbia.edu.

REFERENCES

1. Stretch R, et al. Effect of boarding on mortality in ICUs. Crit Care Med. 2018;46:525–531.
2. Lott JP, et al. Critical illness outcomes in specialty versus general intensive care units. Am J Respir Crit Care Med. 2009 Apr 15;179:676–683.
3. Singer AJ, et al. The association between length of emergency department boarding and mortality. Acad Emerg Med. 2011;18:1324–1329.
4. Al-Qahtani S, et al. The association of duration of boarding in the emergency room and the outcome of patients admitted to the intensive care unit. BMC Emerg Med. 2017 Nov 9;17:34.
5. Rutherford PA, et al. Achieving Hospital-wide Patient Flow. 2nd ed. Cambridge, MA: Institute for Healthcare Improvement, 2020. IHI White PaperAccessed Mar 4, 2021 http://www.ihi.org/resources/Pages/IHIWhitePapers/Achieving-Hospital-wide-Patient-Flow.aspx .
6. Trzeciak S, Rivers EP. Emergency department overcrowding in the United States: an emerging threat to patient safety and public health. Emerg Med J. 2003;20:402–405.

7. McGowan JE, et al. Operating room efficiency and hospital capacity: factors affecting operating room use during maximum hospital census. J Am Coll Surg. 2007;204:865–871.
8. Hall R, et al. Modeling patient flows through the healthcare system. In: Hall RH, editor. Patient Flow: Reducing Delay in Healthcare Delivery. New York: Springer. p. 1–44.
9. Asplin BR, Magid DJ. If you want to fix crowding, start by fixing your hospital. Ann Emerg Med. 2007;49:273–274.
10. Shen Y-C, Hsia RY. Association between ambulance diversion and survival among patients with acute myocardial infarction. JAMA. 2011 Jun 15;305:2440–2447.
11. Bagust A, Place M, Posnett JW. Dynamics of bed use in accommodating emergency admissions: stochastic simulation model. BMJ. 1999 Jul 17;319:155–158.
12. Bain CA, et al. Myths of ideal hospital occupancy. Med J Aust. 2010 Jan 4;192:42–43.
13. Fatovich DM, Hughes G, McCarthy SM. Access block: it's all about available beds. Med J Aust. 2009 Apr 6;190:362–363.
14. Cameron PA, Joseph AP, McCarthy SM. Access block can be managed. Med J Aust. 2009 Apr 6;190:364–368.
15. Forero R, McCarthy S, Hillman K. Access block and emergency department overcrowding. In: Vincent J-L, editor. Annual Update in Intensive Care and Emergency Medicine 2011. New York: Springer. p. 720–728.
16. Jones R. Optimum bed occupancy in psychiatric hospitals. Psychiatry On-line. Epub. 2013 Jul.
17. Green LV, Nguyen V. Strategies for cutting hospital beds: the impact on patient service. Health Serv Res. 2001;36:421–442.
18. Hillier DF, et al. The effect of hospital bed occupancy on throughput in the pediatric emergency department. Ann Emerg Med. 2009;53:767–776 e3.
19. Devapriya P, et al. StratBAM: a discrete-event simulation model to support strategic hospital bed capacity decisions. J Med Syst. 2015;39:130.
20. Dong J, Perry O. Queueing models for patient-flow dynamics in inpatient wards. Oper Res. 2020;68:250–275.
21. Buhaug H. Long waiting lists in hospitals: operational research needs to be used more often and may provide answers. BMJ. 2002 Feb 2;324:252–253.
22. Gorunescu F, McClean SI, Millard PH. A queueing model for bed-occupancy management and planning of hospitals. J Oper Res Soc. 2002;53:19–24.
23. Young T, et al. Using industrial processes to improve patient care. BMJ. 2004 Jan 17;328:162–164.
24. McManus ML, et al. Queuing theory accurately models the need for critical care resources. Anesthesiology. 2004;100:1271–1276.
25. Bitran GR, Tirupati D. Tradeoff curves, targeting and balancing in manufacturing queueing networks. Oper Res. 1989;37:547–564.
26. Koole G, Mandelbaum A. Queueing models of call centers: an introduction. Ann Oper Res. 2002;113:41–59.
27. Alfa AS. Queueing Theory for Telecommunications: Discrete Time Modelling of a Single Node System. New York: Springer, 2010.
28. Armony M, et al. On patient flow in hospitals: a data-based queueing-science perspective. Stochastic Systems. 2015;5:146–194.
29. Jacobson SH, Hall SN, Swisher JR. Discrete-event simulation of health care systems. In: Hall RH, editor. Patient Flow: Reducing Delay in Healthcare Delivery. New York: Springer. p. 211–252.
30. Günal MM, Pidd M. Discrete event simulation for performance modelling in health care: a review of the literature. J Simul. 2010;4:42–51.
31. Dong J, et al. Off-service placement in inpatient ward network: Resource pooling versus service slowdown. Accessed March 18, 2021. http://www.columbia.edu/~jd2736/publication/Offservice.pdf.
32. GitHub. Implementation of a Patient Flow Model to Optimize Inpatient Hospital Capacity. Hu Y, et al. Accessed Mar 4, 2021. https://github.com/YueHu-CU/Implementation-of-a-patient-flow-model-to-optimize-inpatient-hospital-capacity.
33. Gillam S. Rising hospital admissions. BMJ. 2010 Feb 2;340:c636.