



Operations Research

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

Using Hospital Admission Predictions at Triage for Improving Patient Length of Stay in Emergency Departments

Wanyi Chen, Nilay Tanik Argon, Tommy Bohrmann, Benjamin Linthicum, Kenneth Lopiano, Abhishek Mehrotra, Debbie Travers, Serhan Ziya

To cite this article:

Wanyi Chen, Nilay Tanik Argon, Tommy Bohrmann, Benjamin Linthicum, Kenneth Lopiano, Abhishek Mehrotra, Debbie Travers, Serhan Ziya (2022) Using Hospital Admission Predictions at Triage for Improving Patient Length of Stay in Emergency Departments. Operations Research

Published online in Articles in Advance 29 Nov 2022

. <https://doi.org/10.1287/opre.2022.2405>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2022, INFORMS

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.







For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

Crosscutting Areas

Using Hospital Admission Predictions at Triage for Improving Patient Length of Stay in Emergency Departments

Wanyi Chen,^a Nilay Tanik Argon,^{b,*} Tommy Bohrmann,^c Benjamin Linthicum,^d Kenneth Lopiano,^e Abhishek Mehrotra,^d Debbie Travers,^f Serhan Ziya^b^aMassachusetts General Hospital, Harvard Medical School, Boston, Massachusetts 02140; ^bDepartment of Statistics and Operations Research, University of North Carolina, Chapel Hill, North Carolina 27599; ^cAnalytical Partners Consulting LLC, Research Triangle Park, North Carolina 27709; ^dSchool of Medicine, University of North Carolina, Chapel Hill, North Carolina 27599; ^eRoundtable Analytics, Inc., Research Triangle Park, North Carolina 27709; ^fSchool of Nursing, Duke University, Durham, North Carolina 27710

*Corresponding author

Contact: wchen38@mg.harvard.edu,  <https://orcid.org/0000-0001-8523-1032> (WC); nilay@email.unc.edu,  <https://orcid.org/0000-0002-6814-0849> (NTA); tommy_bohrmann@med.unc.edu (TB); benjamin_linthicum@med.unc.edu,  <https://orcid.org/0000-0003-4322-7662> (BL); kenny@roundtableanalytics.com (KL); abhi_mehrotra@med.unc.edu,  <https://orcid.org/0000-0002-9119-831X> (AM); debbie.travers@duke.edu,  <https://orcid.org/0000-0002-5656-8610> (DT); ziya@unc.edu,  <https://orcid.org/0000-0003-1558-6051> (SZ)**Received:** January 30, 2020**Revised:** June 23, 2021; June 23, 2022**Accepted:** September 29, 2022**Published Online in Articles in Advance:** November 29, 2022**Area of Review:** Policy Modeling and Public Sector OR<https://doi.org/10.1287/opre.2022.2405>**Copyright:** © 2022 INFORMS**Abstract.** Long boarding times have long been recognized as one of the main reasons behind emergency department (ED) crowding. One of the suggestions made in the literature to reduce boarding times was to predict, at the time of triage, whether a patient will eventually be admitted to the hospital and if the prediction turns out to be “admit,” start preparations for the patient’s transfer to the main hospital early in the ED visit. However, there has been no systematic effort in developing a method to help determine whether an estimate for the probability of admit would be considered high enough to request a bed early, whether this determination should depend on ED census, and what the potential benefits of adopting such a policy would be. This paper aims to help fill this gap. The methodology we propose estimates hospital admission probabilities using standard logistic regression techniques. To determine whether a given probability of admission is high enough to qualify a bed request early, we develop and analyze two mathematical decision models. Both models are simplified representations and thus, do not lead to directly implementable policies. However, building on the solutions to these simple models, we propose two policies that can be used in practice. Then, using data from an academic hospital ED in the southeastern United States, we develop a simulation model, investigate the potential benefits of adopting the two policies, and compare their performances with that under a simple benchmark policy. We find that both policies can bring modest to substantial benefits, with the state-dependent policy outperforming the state-independent one particularly under conditions when the ED experiences more than usual levels of patient demand.**Funding:** This work was supported by the National Science Foundation [Grants CMMI-1234212 and CMMI-1635574].**Supplemental Material:** The online appendix is available at <https://doi.org/10.1287/opre.2022.2405>.**Keywords:** patient flow • healthcare operations • queuing • Markov decision processes

1. Introduction

Long waits and congestion at emergency departments (EDs) have long been recognized as a challenging problem to tackle. EDs have tested with and in some cases, adopted various novel operational methods to alleviate congestion levels and generally improve ED throughput. However, there seems to be a general agreement that substantial improvements can only be achieved with a systems-level perspective by adoption of policies that recognize and exploit the fact that ED operations are closely tied to the operations and decisions in the main hospital. Arguably, the close relationship between the

ED and the main hospital and the potential benefits of effective coordination between the two can be seen best through patient boarding times,¹ which constitute one of the major components of patients’ ED length of stay (LOS).² When there are hospital beds available, a decision to admit a patient from the ED to the main hospital initiates what we call the transfer preparation process (TPP), which includes identification of a specific bed and a care team for the patient in the main hospital and carrying out all the essential tasks for the patient’s transfer from the ED to the main hospital. When there are no hospital beds immediately available but a decision is made

to admit an ED patient, the TPP for the patient starts only after a hospital bed for the patient becomes available. Thus, the TPP for a patient and any wait for a bed to become available, both of which can take a significant amount of time, are essentially what determine the patient's boarding time. There are many reasons as to why the TPP time for a patient can take long. Armony et al. (2015) list 13 possible factors (each related to equipment availability, staff availability, ED-hospital synchronization, or other issues related to the hospital practices) that add to the duration of TPP for a given patient.

In many respects, a patient boarding in the ED is a sign of inefficient use of resources; the patient, who no longer needs ED care, continues to occupy an ED bed and demands the attention of ED staff, both of which could in fact be used for other patients either in the ED treatment area or in the waiting room. It is thus clear that keeping boarding times as short as possible is highly important, and what the ED can do to achieve this goal without active participation of the main hospital is very limited. The objective of this paper is to develop a framework that encourages and demands active involvement of the ED and the main hospital in an effort to reduce boarding times and investigate the potential benefits of adopting this framework in practice.

The framework we propose is based on the following simple idea; at triage, identify the patients who have a good chance to be admitted to the hospital at the end of their ED stay and request a bed for those patients right away at triage without waiting for the eventual disposition decision for those patients. We call this practice *bed request at triage* (BeRT). If a hospital bed is available for the patient at the time of BeRT, TPP for the patient is initiated right away; otherwise, the request is kept in a queue until a hospital bed for the patient becomes available, at which time TPP is initiated. BeRT essentially aims to parallelize, to the extent possible, two main tasks performed for a patient while the patient is in the ED: the emergency care provided to the patient and TPP, which according to the current practice, are performed in sequence, one after the other.

The potential benefit of the parallelization achieved by BeRT is that the patient for whom a bed is requested at the time of triage will, once she is admitted, board in the ED for an amount of time that is much shorter than she would under the current system because the main hospital will be ready for the patient's transfer either by the time of the patient's admit decision or soon after that. The downside is that the prediction at triage might be incorrect, and the patient might ultimately end up not being admitted to the hospital. This would be a problem because that would mean that hospital resources were not used in an efficient manner (i.e., resources that were used for TPP could have been used for a more urgent or essential task). In addition, this would have a negative impact on the future potential benefits of the practice,

particularly during the early stages of implementing such a framework because the hospital staff involved might grow doubtful that the bed requests made by the ED will indeed result in actual patient transfers and get increasingly more reluctant to act urgently based on prediction alone without a definite admit decision.

One way an ED can identify patients with high likelihood of admission is by looking at patients' emergency severity index (ESI) triage acuity level, which strongly correlates with patients' eventual disposition decision, with lower ESI levels being closely associated with a higher likelihood of admission. For instance, an ED can decide to request beds early for all ESI-1 and ESI-2 patients, at least during certain times of the day, believing that their likelihood of admission is sufficiently high. (Including ESI-3 patients would not work well, as we discuss in more detail in Section 6 because for most EDs, that would mean very large numbers of false early bed requests every day.) Such a policy would be immediately implementable without any further development in any ED that uses ESI classification for triage. EDs that use alternative triage schemes can also adopt similar policies as long as acuity classes correlate with hospital admissions. However, it is not clear whether the simple criteria such a policy applies for requesting beds early would work well. ESI-1 and ESI-2 patients might indeed be the patients who are generally more likely to be admitted, but it is not clear whether beds should be requested early for all such patients. The ED might want to be more selective and request beds early for only a subset of ESI-1 and ESI-2 patients by identifying those who are particularly more likely to be admitted to the hospital. Furthermore, when deciding whether the likelihood of admission is high enough for an early bed request, the ED might want to take into account operating conditions in the ED, such as the crowding level. The framework we develop in this paper can be used to help determine specifically for which patients beds should be requested early, possibly depending on the ED census.

In our framework, we make the decision of whether a hospital bed should be requested for a patient at the time of triage in two steps. First, for each arriving patient classified as ESI-1 or ESI-2, based on the information available only at the completion of triage, such as patient complaints, triage acuity level, demographic information, etc., we estimate the probability that the patient will eventually be admitted to the hospital. Then, based on this prediction and possibly some relevant system-level information, such as ED census, we make a decision as to whether a bed for the patient should be requested. This paper is mainly about this second step. Specifically, we are interested in the following question. Given the probability that a patient will be admitted to the main hospital based on what we know at the time of triage, should we place an early bed request for the patient? (We only consider ESI-1 and ESI-2 patients because in

the ED where our data came from, patients in these two groups constitute a sufficiently large pool of patients with a high probability of admission. Furthermore, our simulation studies showed that broadening this pool would only hurt the performance measures of interest. Note, however, that one can use this framework without restricting the attention to ESI-1 and ESI-2 patients alone, or one could consider a smaller or larger subset of the patients.) The first question of how to estimate the admission probability at triage is presented elsewhere in Mehrotra et al. (2017) and is out of our scope here. However, it might be helpful to note that this estimation is done through logistic regression with predictor variables, including patient ESI, patient age, and the presence or absence of chief complaints that are highly associated with hospital admission, such as respiratory distress. This regression-based probability estimation tool is named the admission prediction tool (APT).

As we discuss in Section 2, we are not the first to propose the idea of placing hospital bed requests early for patients who are predicted to be eventually admitted to the main hospital. To the best of our knowledge, however, no prior work has developed a complete framework that prescribes not only how admission predictions should be made but also, how exactly these predictions should be used to make early bed requests from the hospital and then, tested the potential benefits of adopting such a framework. With this paper, we take the first step toward filling this gap. When developing our framework, we used data from an academic hospital ED in the southeastern United States. We have also used the same ED as the setting for our simulation study and for the pilot study we conducted to investigate the feasibility of and the challenges associated with the practice of making early bed requests. However, it is important to note that, as it should be clear from the rest of this paper, the proposed methodology is highly general and can easily be adapted to other EDs and hospitals, particularly those in the United States.

In what follows, after a literature review (Section 2), we first give an overview of three early bed request policies we will analyze in this paper (Section 3). Of these three policies, the *emergency severity index-based policy* (ESIB) is the simplest and can be used without any mathematical or statistical model development or analysis, and therefore, it will serve as a benchmark for the other two policies, namely the *fixed threshold policy* (FT) and the *census- and time-based threshold policy* (CTT), which we develop and propose in this paper. Then, we give a short description of APT and the data set we used throughout the paper (Section 4). Section 5 is mainly devoted to the development of CTT. However, to motivate the mathematical decision problem, which CTT is based on, it starts with discussing the fundamental challenges and the basic issues that need to be considered when making early bed requests. Essentially, we argue that the overall goal should be to reduce average patient length of stay

by making early bed requests but without leading to frequent false bed requests. Then, we develop mathematical models that capture this basic trade-off, analyze these models, identify optimal or “good” solutions, and then, based on these solutions, provide a description of CTT, which prescribes whether an early bed should be requested for a patient given the patient’s admission probability and the ED census level. We also provide a formal description of FT, which prescribes actions independently of the ED census level. Then, in Section 6, we report the results of a thorough simulation study we conducted to investigate the performances of the policies we propose. Our findings strongly suggest that the improvements in patient length of stay that would be achieved by adoption of the policies we propose, even with conservative levels of false bed requests, can be significant, particularly when the ED experiences high levels of patient load. The benchmark policy ESIB also performs well; however, the performances of FT and CTT are both statistically superior. We also find that CTT always performs at least as well as FT while outperforming it in most of the scenarios considered. Finally, in Section 7, we provide our concluding remarks. Proofs of our analytical results and details of our simulation study are presented in the online appendix.

2. Relevant Literature

Several papers from the emergency medicine literature have investigated how hospital admissions can be predicted in advance and discussed ways of using admission probability estimates for making better decisions and improving patient flow. For example, see Boyle et al. (2012), Peck et al. (2012, 2013), Crilly et al. (2015), and Somanchi et al. (2022). To the best of our knowledge, however, Qiu et al. (2015) is the only paper to date that makes a specific proposal as to how the probability estimates can be used, and it is important to note that even in this paper, what the authors are mainly interested in and their general approach to their problem are completely different from ours. Specifically, Qiu et al. (2015) is mainly interested in determining the “optimal” time for requesting a bed for a patient whose admission probability is known. In contrast, we assume that if a bed is to be requested early for a patient, it has to be requested immediately after triage, but we focus on the question of whether an early bed request should be made in the first place considering the operational implications of the decision not only on that particular patient alone but at the system level on the collection of all the patients.

Although we are not aware of any prior work on the decision problem we investigate in this paper, it is important to note that over the last several years, the operations community has shown increasingly more interest in problems related to ED and hospital operations, with some focusing specifically on patient flow.

Here, we review those that appear to be closest to our work. (For a recent review of modeling and analytical work on patient flow management, see Dai and Shi (2021).) Using data from a Singapore hospital, Shi et al. (2016) study in-hospital operations with the goal of reducing ED boarding times. Thus, the paper, considering this general objective, is similar to ours as we also aim to reduce ED length of stay through shortening of ED boarding times. However, the two papers are fundamentally different in how they envision that goal to be achieved. Shi et al. (2016) propose and test, via simulation, a hypothetical hospital discharge policy that pushes the discharge times to earlier in the day so that more in-hospital beds are available by the time the bulk of the ED patients arrive. In contrast, the policies we propose and test in this paper call for requesting hospital beds earlier, at the time of triage, for patients who have a high probability of being admitted.

Saghafian et al. (2012, 2014) investigate the potential benefits of streaming patients in the ED. Although the former paper mainly considers streaming with respect to whether a patient is likely to be admitted to the hospital or discharged from the ED, the latter considers streaming with respect to patient complexity. Armony et al. (2015) provide a detailed mathematical and statistical description of the operational features of an ED located in Israel, outline the fundamental blocks of an ED-hospital system, and provide directions for future research. Huang et al. (2015) view the patient prioritization problem in an ED as a queueing control problem and develop policies that help make the decision of which patient to see next in the ED. Xuang and Chan (2016) develop policies for patient admission to an ED taking into account information on future arrivals to the ED, with the goal of reducing waiting times. Chan et al. (2017a) are interested in the question of when to perform patient inspections for the purpose of determining whether they can be discharged from the hospital. Kamali et al. (2019) investigate the question of under what conditions it would be beneficial to have a physician responsible for patient triage in addition to a nurse.

Ang et al. (2016) develop a method called Q-Lasso for predicting ED waiting times, implement the method in an actual ED, and find that the prediction error with Q-Lasso is significantly smaller than that under the best-performing rolling average policy, a type of estimation method that is currently in use by many hospitals. Chan et al. (2017b) study the impact of ED boarding on patients' intensive care unit length of stay, develop a queueing model that captures the phenomenon of service times being negatively affected by delays in access to service, and propose an approximation for the expected work in the system. Motivated by ED admissions to the main hospital, Dai and Shi (2017) develop a two-timescale queueing model, with the two timescales being an important feature that helps capture the length of stay

in terms of hours and days, essentially enabling a more realistic formulation of hospital admissions from the ED. The authors then provide expressions for the steady-state distribution for the number of patients in the hospital at midnight and several performance measures. Dong et al. (2019) study capacity management questions for inpatient wards, including strategies like off-service placement, which the hospitals typically employ in response to high patient demand levels. The paper provides solutions for and insights into how patient demand-bed supply balance can be achieved. In order to capture the relationship between early discharges and increased readmission risk, Shi et al. (2021) develop a Markov decision process (MDP) formulation with the goal of determining how many and which patients to discharge on each day given predictions of readmission at the individual patient level, develop a heuristic solution, and then, investigate the potential benefits of using this solution when making discharge decisions.

There are also papers that carry out empirical analysis, with the objective of developing a better understanding of and providing insights into managing ED and hospital operations. For some recent examples of this line of work, see Kuntz et al. (2015), Song et al. (2015), Batt and Terwiesch (2016), Diwas and Terwiesch (2017), Long and Mathews (2017), and Ding et al. (2019).

Finally, it is important to note that the mathematical model we use to develop policies for making early bed request decisions is a type of dynamic queueing control problem, similar to that of Huang et al. (2015), and thus, our paper can be seen as a contribution to that area as well. However, although there are many papers written on the topic of how to dynamically assign servers to different tasks in a stochastic network (see, e.g., Andradottir et al. 2003, Zayas-Caban et al. 2016, Legros et al. 2018), we are not aware of any papers on the question of how to dynamically determine what kind of service to perform on a given job depending on the system state, which is essentially the decision our mathematical models deal with.

3. From Simple to Complex, Three Different Approaches to BeRT

We are not aware of any emergency department that has a policy of requesting hospital beds for any of their patients prior to their disposition decisions. It will be reasonable to assume, however, that if an early bed request policy was to be implemented at an emergency department, it would try to identify patients who are highly likely to be admitted to the hospital and request beds early for those. One way to identify those patients is through their ESI classifications. The ESI class for each patient is readily known at the time of triage and is strongly correlated with admissions, with patients at lower levels being more likely to be admitted (see, e.g.,

Chen et al. 2020). Thus, without going through the trouble of developing an estimation model for the probability of admission, an ED can simply choose to implement a policy that requests beds early according to patients' ESI levels. For example, the policy might call for requesting hospital beds early for all the patients classified as ESI-1 or ESI-2, at least during certain times of the day. In this paper, we call this policy ESIB. (Clearly, the ED might instead choose to request beds early only for ESI-1 patients or perhaps, include ESI-3 patients as well. However, as also confirmed by our simulation study (see Section 6.1), those policies would not work well because excluding ESI-2 patients will mean beds will be requested early only for a very small number of patients, whereas including ESI-3 patients will mean beds will be requested early for such a large segment of the patient population that there will be many false bed requests.)

ESIB is an easily implementable policy, but it is not flexible. Requesting beds early for all ESI-1 and ESI-2 patients may not work all that well, and thus, it might be reasonable to identify ESI-1 and ESI-2 patients who are particularly more likely to be admitted and request beds early only for those patients. One way to do this is by developing a model that can be used to estimate the probability that a given ESI-1 or ESI-2 patient will be admitted to the hospital, use this model to identify those patients who are substantially more likely to be admitted based on some probability threshold, and request beds early only for those patients. We call this policy FT. Note that FT requests beds early for a subset of the patients for whom early bed requests would be made under ESIB, meaning that, just like ESIB, it restricts these requests to ESI-1 and ESI-2 patients, but for simplicity, we do not explicitly highlight this in the name of the policy.

With a carefully chosen threshold, FT would likely improve upon ESIB. However, the improvement might be even higher if the threshold is changed dynamically depending on time of day and ED census level. For example, when ED is crowded and waiting times are long, the limited bed capacity would be even more valuable, and thus, it might make sense for the ED to take a little more risk and set a lower threshold. In general, however, it is not clear precisely how one should change the threshold level with changing the ED census. The core of this paper, presented in Section 5, is devoted to that question. Our analysis in that section leads to a dynamic policy, which we call CTT. Note that just like FT, CTT also restricts early bed requests to ESI-1 and ESI-2 patients.

ESIB, FT, and CTT can be seen as going from simple to complex, with CTT being the most sophisticated of the three. As we noted, ESIB does not require any new model development and can easily be implemented in practice. Therefore, in Section 6, we will use it as a benchmark policy in our performance analysis of the policies we propose (i.e., FT and CTT). Both FT and CTT need a

tool for estimating the probability of admission for patients based on information available at the time of triage. Next, in Section 4, we describe this tool, which we call the APT, along with the data used in its development and the rest of the paper.

4. Description of the Data and the APT

We considered a data set of patient visits to an academic hospital ED in the southeastern United States collected during the year 2012. According to the data, the ED had 67,203 visits during the year, corresponding to about 184 patients per day on average. Roughly, about 29.6% of the patients were admitted to the hospital. The average ED length of stay was about 356 minutes, the average ED workup time (time between rooming until the disposition decision) was 212 minutes, and the average boarding time was 235 minutes. After a cleanup of the data set, we ended up with approximately 65,065 patient entries. Each patient entry contained detailed data that included the following: time stamps (arrival time, disposition decision time, departure time, etc.), chief complaints, triage acuity/ESI, and demographic information (age, gender, race, etc.). We considered alternative logistic regression models, which can be used to estimate the hospital admission probability for ED patients based on information only available at the time of triage. Complete details of this analysis were presented in Mehrotra et al. (2017), but here, we give a summary of our basic findings and a rough description of the model chosen as the best in the end (i.e., what we call APT in this paper).

The key predictors of admission turned out to be ESI levels 1–3, the three most urgent triage levels of five; age groups 55–70 and above 70; and the existence of certain chief complaints, such as respiratory distress. Thus, the estimate for the admission probability as determined by our model, APT, is a function of the patient's ESI level, age, and description of his or her main complaints at the time of triage. Hence, APT is basically a simple mathematical expression that returns an estimate for the probability of hospital admission for a patient given the presence or absence of these predictors.

Our goal is to use the admission probability estimates to determine whether a hospital bed for a patient should be requested at the time of triage. A reasonable way of making this decision is by comparing the admission probability estimate with some threshold level ξ , which may or may not be fixed at all times, and requesting a bed if the estimate is large enough (i.e., larger than ξ). In this paper, we are concerned with how this threshold level ξ should be determined, possibly depending on system conditions, not the performance of APT itself. The policies we propose would work regardless of the admission probability estimator used. Nevertheless, it is important to note that the predictive power of APT is reasonably good. For example, if ξ is set to 0.9 at all

times, approximately 91% of early bed requests will result in actual patient admits to the hospital based on our data set (assuming that early bed requests are made for patients at any ESI level). The percentage drops only to 86 if ξ is set to 0.8. However, the reader should also note that the data set we used when developing APT was rather limited, and we believe that there is definite room for improvement with a more comprehensive study that involves a richer data set.

5. CTT and FT

Requesting beds early for patients who have a good chance to be admitted to the hospital has clear benefits for ED operations (at least from the perspective of reducing overall ED patient length of stay), and thus, one might question why not adopt this practice widely and request hospital beds early even for patients who have a small but nonnegligible probability of being admitted to the hospital. However, looking at the issue from an ED perspective alone would be unhelpful as the practice would clearly require significant involvement of the main hospital staff as well. From the hospital side, problems would arise if the ED frequently demands a bed from the main hospital only to cancel the request later on after realizing that the patient does not need to be admitted to the hospital after all. There are mainly two reasons why this would be problematic. First, a false bed request would mean that the hospital staff spent their time on a set of tasks that in fact did not need to be done, at least right away. Second, false requests, especially early on in the implementation of the policy of requesting beds at triage, are likely to present a significant impediment to the policy's adoption and it being embraced by the hospital staff. They could possibly lead to the hospital staff losing trust in the policy, making them reluctant to respond to future early bed requests and leading to friction between the ED and the main hospital staff.

In short, there are two competing goals one needs to keep in mind: reducing the overall ED length of stay and keeping false bed requests at minimum. Obviously, one can easily keep false bed requests at minimum, in fact at zero, by simply not making any early bed requests. On the other hand, the reduction in ED length of stay would be maximized (at least in theory) if early bed requests are made as frequently as possible. These are two extreme positions one can choose to take. The former essentially describes the status quo, and we would like to improve upon that; the latter is simply not practically feasible because frequent false bed requests would very likely lead to the collapse of the whole policy of requesting beds early soon after its adoption. Thus, the ED and the hospital would prefer to operate somewhere in between, but it is not clear exactly where to operate and also, how to get there.

One way to balance the two competing goals described is by simply agreeing on a certain level of incorrect early

bed requests that the ED and the hospital would be willing to bear and set a threshold level on the admission probabilities so that when early bed requests are made for all the patients whose admission probabilities exceed this threshold level, the target incorrect request level is met. For example, the hospital might decide that it would be willing to live with up to two incorrect requests per day on average. Then, using the empirical distribution for the admission probabilities as well as the estimates on the daily number of patient arrivals, one can compute what threshold level would lead to the expected number of incorrect requests per day being equal to two and then, implement a policy that will request beds early for all the patients with admission probability that is larger than this threshold. However, one might argue that long boarding times would be particularly detrimental when ED is crowded, and thus, it might make sense to set the threshold level dynamically depending on system conditions. The hospital might still want to meet a certain level in regard to the overall frequency of incorrect requests, but it might find it more preferable to live with potentially higher levels of false requests (by decreasing the threshold) when the ED is crowded and lower levels of false requests (by increasing the threshold) when the ED is not crowded. The question then is how to set the threshold level that will trigger early bed requests dynamically depending on ED crowding levels.

To answer this question, we use the following approach; we first cast the decision problem described from a perspective of cost minimization, with costs tied to each unit of time a patient spends in the ED and each false bed request. We do not view these costs in dollar terms but as penalties that help define our objective in a convenient fashion, at least in some approximate way. One additional advantage of casting the problem as cost minimization is that, as we shall see later in the paper, the cost parameters can be used as tuning parameters at the policy implementation stage. Even under the cost minimization objective, however, it is not clear how exactly one should formulate the problem. One direction would be to consider a highly complex dynamic program that captures the actual ED/hospital system as realistically as possible and develop an approximate solution method. This would, however, lead to solutions, if possible at all, that are difficult for the practitioners to interpret and make adjustments in practice. Another direction would be to consider the analysis of simplified formulations that capture the basic dynamics only, generate insights into the type of policies that might work well in practice, build on the solutions to these simple models, and propose practical policies. In this paper, we follow this latter approach.

To be clear, our ultimate objective remains the somewhat vague but practically more useful goal of reducing the overall length of stay while keeping false bed requests low. We will come back to this objective later in the paper, in Section 6, where we make policy comparisons, but in the

rest of this section, our main focus will be on the development of cost models, their mathematical analyses, and the development of practical policies based on these analyses.

As we stated, our solution approach will be through the analysis of simplified mathematical formulations. However, first, it will be helpful to provide a formal description of the problem with its close to full complexity so that the reader can better understand in what way our simplified models fit. As we explain, even for this more complex formulation, we will need to make some assumptions that do not hold in practice. However, it is important to note that the policies we will be proposing in the end are not tied to these assumptions and that they will be tested using a realistic simulation model of an actual ED.

5.1. Problem Description

The problem we are interested in is a dynamic decision problem. Specifically, our objective is to minimize the total long-run average cost of keeping patients waiting and making incorrect early bed requests by deciding, for each patient, at the time of triage whether a hospital bed for the patient should be requested given the patient's probability of admission to the hospital and the system state. What exactly this system state should include would depend on the degree of simplification one is willing to make through various modeling assumptions. In this section, our goal is to keep the formulation as general as possible but only resort to some simplifying assumptions when their absence would lead to significant modeling and notational complexity. As we explain, even under these simplifying assumptions, the resulting MDP formulation would still be too complex that its complete description would be long and heavy in notation. Therefore, in this section, we provide a general outline for this formulation mainly by stating the assumptions needed and formally describing the state space.

We consider a service system where patients go through one or two phases depending on their disposition decision. The first phase corresponds to the ED stay, and the second phase corresponds to the hospital stay. There are K_{ED} servers for the first phase (corresponding to the ED beds) and K_H servers for the second phase (corresponding to hospital beds). There are 10 types of patients corresponding to different combinations of ESI levels (five levels) and age (two levels: pediatric and adults). We assume that patients of type $i = 1, 2, \dots, 10$ arrive according to a Poisson process with rate λ_i ($0 < \lambda_i < \infty$). As soon as a patient arrives, the patient goes through triage. We assume that the time it takes to perform triage on a patient is negligible, and therefore, the probability of admission for the patient is known as soon as the patient arrives at the ED. Let Z_k^i denote the random variable representing the probability that the k th type i patient to arrive will be admitted to the hospital. We assume that for $i = 1, 2, \dots, 10$, $\{Z_k^i\}_{k=1}^\infty$ is a sequence of independent and identically distributed (iid) random variables with

the common probability mass function specified as $P\{Z_k^i = \alpha_j\} = q_j^i$ for $\alpha_j \in \Omega$ and $k \in \{1, 2, \dots\}$, where $\Omega = \{\alpha_1, \alpha_2, \dots, \alpha_M\}$ is the finite set of possible values Z_k^i can take.

Once the triage for the patient is complete, if there is an ED bed available, the patient starts the first-phase service. If there are no beds available, the patient starts waiting for an ED bed. Waiting patients are accepted into first-phase service as ED beds become available according to the order determined by their ESI levels and arrival times. Patients with lower ESI levels have priority over those with higher ESI levels. Within the same ESI level, the order is according to first come, first served (FCFS). The first-phase service basically consists of what we call the *ED workup time*, the time between the patient's rooming in the ED and the time a disposition decision for the patient is made, followed by either the *boarding time* or *discharge time* depending on whether the patient is admitted to the hospital or discharged from the ED.

The boarding time depends on whether a bed has been requested for the patient at the time of triage and whether a bed was available and assigned to the patient at the time the request was made. If a bed has not been requested at the time of triage, then the request is made at the time the patient is admitted to the hospital, in which case the boarding time is assumed to be equal to the TPP time if a hospital bed is available for the patient. If a bed has already been requested at the time of triage and the bed was available, then the boarding time is assumed to be equal to the time remaining until TPP is over, which might possibly be equal to zero. If, however, regardless of when a hospital bed is requested, there are no hospital beds available at the time of bed request, then the bed request joins a queue waiting for a hospital bed to become available. When a new hospital bed becomes available for allocation, the bed is assigned to one of the requests in the queue randomly. In that case, boarding time lasts until a bed is allocated to the patient and TPP, which starts right after allocation, is complete. If a patient is not admitted to the hospital, the patient continues to occupy the ED bed until the discharge process is complete. If a hospital bed was allocated to the discharged patient at the time of the patient's triage, that bed becomes available for other patients. If a bed was not allocated but a request was waiting in a queue, the request is canceled. We assume that ED workup times, TPP times, and discharge times each are a separate sequence of iid exponentially distributed random variables.

Once the first-phase service of the patient is over, the patient moves to the second-phase service (hospital stay). The second-phase service times (i.e., times spent in hospital beds) are also assumed to be iid exponentially distributed random variables.

Assuming that it costs the system c_w for each unit of time a single patient spends in the ED, c_{tr} to request a hospital bed at the time of triage, and c_{ad} to request a hospital bed after disposition, the problem can be

formulated as an MDP with the objective of minimizing the long-run average cost. Decision epochs are the times at which patients are admitted to the ED. Specifically, every time a patient starts her first-phase service, a decision is made as to whether a hospital bed should be requested for the patient in advance. The system state can be described as a vector with a size that depends on the state itself. Specifically, the number of dimensions needed to represent the state vector would depend on N , the number of patients in the ED. To completely describe system states, we introduce the following notation. For $j = 1, 2, \dots, N$, let R_j denote the current state for patient j who is the j th patient to have arrived at the ED among the N patients currently in the ED:

$$R_j = \begin{cases} 1 & \text{if patient } j \text{ is waiting for an ED bed} \\ 2 & \text{if patient } j \text{ is going through ED workup,} \\ & \text{a hospital bed has been requested for the patient,} \\ & \text{a hospital bed has been allocated but is not ready} \\ 3 & \text{if patient } j \text{ is going through ED workup,} \\ & \text{a hospital bed has been requested for the patient} \\ & \text{but has not been allocated} \\ 4 & \text{if patient } j \text{ is going through ED workup,} \\ & \text{a hospital bed has been requested for the patient,} \\ & \text{and the bed is ready} \\ 5 & \text{if patient } j \text{ is waiting to be discharged} \\ 6 & \text{if patient } j \text{ is boarding, a hospital bed has been} \\ & \text{allocated but is not ready} \\ 7 & \text{if patient } j \text{ is boarding, a hospital bed has been} \\ & \text{requested but has not been allocated.} \end{cases}$$

Let m_j denote the hospital admission probability for patient j , which has to be part of the system state because this information is needed at the disposition time for patient j to determine whether the patient will be admitted to the hospital.

To model the bed occupancies at the hospital, we let H_1 denote the number of occupied hospital beds and H_2 denote the number of hospital beds available for allocation. Note that at any given time, in addition to the occupied beds and the beds available for allocation, there can also be hospital beds that have been allocated but with TPP in progress and also, hospital beds that have been allocated for which TPP is complete but the patient to be admitted is still going through ED workup. However, we do not need to keep track of these beds separately because they can be determined through R_j . Then, the system state can be described by the vector $(N, R_1, R_2, \dots, R_N, m_1, m_2, \dots, m_N, H_1, H_2)$, and the state space can be expressed as

$$\mathbb{X} = \left\{ (n, x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_n, h_1, h_2) \mid n \in \mathbb{Z}; \right. \\ \left. x_i \in \{1, 2, 3, 4, 5, 6, 7\}; \right. \\ \left. y_i \in \Omega \text{ for } i = 1, 2, \dots, n; h_1, h_2 \in \mathbb{Z} \text{ and } h_1 + h_2 \right. \\ \left. + \sum_{i=1}^n \mathbf{1}_{\{x_i \in \{2,4,6\}\}} = K_H \right\}.$$

Under the assumptions we stated in this section, we can model this problem as an MDP. In particular, given the state representation, one can write the transition probabilities and express the long-run average optimality equations. However, doing that would only be an exercise in modeling without benefits because it is clear that providing a direct solution to this problem with no further simplification would be extremely challenging. (To get an idea about why transition probabilities would be somewhat complex, consider an event that completes the boarding time of a patient. This single event could possibly lead to a change in the values of every component of the state vector with the exception of H_2 because some of the indices of the other patients in the ED will need to be updated as well.) It is also important to note that despite the complexity of this MDP, it would still fail to incorporate some of the basic features of the actual system. For example, we know that patient arrival rates very much depend on the day of week and time of day, and times spent in the ED and hospital are not exponentially distributed. It would also make much more sense to assume that newly vacated hospital beds are allocated to outstanding requests in an FCFS fashion, not randomly as we assumed. Relaxing any one of these assumptions, however, would make the problem substantially more complex.

In short, attempts to model this system realistically and developing methods that solve this problem directly do not appear to be a promising path. Therefore, our goal is to approach this problem from the opposite angle by considering highly simplified formulations but using their analysis to devise methods that can be implemented under realistic conditions. Next, in Section 5.2, we start following this path by focusing on a single patient/bed in isolation.

5.2. A Simple Decision Model for a Single Patient/Bed in Isolation

Consider an ED patient whose probability of admission to the hospital, as computed at the completion of triage, is z . Let S denote the generic random variable that represents the ED workup time. Let B denote the random variable representing the TPP time for the patient. We assume that if a hospital bed is requested for the patient, a hospital bed can be assigned to the patient right away so that B , the TPP time, is also equal to the time between the placement of a bed request and the time the hospital and the ED are ready for the patient's transfer from the ED to the inpatient bed. We also assume that B does not depend on whether the bed is requested early at the completion of triage or later after the disposition decision is made. Given that, it will be reasonable to assume that if the patient is discharged from the ED, then the ED bed will be occupied for S units of time in total. If a hospital bed is not requested early for the patient but the

patient ends up being admitted to the hospital, then the ED bed will be occupied for $S + B$ units of time; if a hospital bed is requested early for the patient and the patient is admitted to the hospital, then the ED bed will be occupied for $\max(S, B)$ units of time. (Given our focus, we assume without loss of generality that the additional time a discharged patient spends in the ED is zero.) For the sake of simplicity, we assume that S and B are independent, S has an exponential distribution with mean $1/\gamma_S$, and B has an exponential distribution with mean $1/\gamma_B$.

The assumptions regarding costs are the same as described in Section 5.1, with c_w denoting the per patient per unit time cost of having a patient in the ED and c_{tr} and c_{ad} denoting the costs of putting in bed requests at the time of triage and after disposition, respectively. (Note that there is no need to consider an additional cost for a false bed request even if $c_{ad} = c_{tr}$, because for patients who are not admitted but for whom a bed was requested at the time of triage, c_{tr} would incur, essentially capturing the cost of a false bed request.)

Then, if a bed is not requested (i.e., TPP is not initiated) at the time of triage, the expected cost is

$$c_w(E[S] + zE[B]) + zc_{ad} = c_w\left(\frac{1}{\gamma_S} + z\frac{1}{\gamma_B}\right) + zc_{ad}.$$

If, however, a bed is requested at the time of triage, then the expected cost is

$$\begin{aligned} & c_w\left(E[S] + zE[\max(B, S) - S]\right) + c_{tr} \\ &= c_w\left(\frac{1}{\gamma_S} + z\frac{1}{\gamma_B} - z\frac{1}{\gamma_S + \gamma_B}\right) + c_{tr}. \end{aligned}$$

One can then show that requesting a bed at the time of triage is at least as good as waiting until the ED workup is over if and only if

$$z \geq \frac{c_{tr}}{c_{ad} + c_w(\gamma_S + \gamma_B)^{-1}}. \quad (5.1)$$

Not surprisingly, the simple single-bed/patient formulation leads to a simple threshold-type decision rule: request a hospital bed at the time of triage if and only if the admission probability for the patient is sufficiently high (i.e., above the threshold for which we have a mathematical expression). The simple nature of the decision rule is appealing, but the reader might have noted two potential limitations and challenges associated with this approach. First, it is not clear how one would set the cost terms c_w , c_{tr} , and c_{ad} . This is an important question and will be addressed later in Section 5.5. However, even if the cost terms can be determined somehow, there is the question of whether it makes sense to use a formulation that assumes a single patient/bed in isolation when, in fact, we know that EDs are typically highly crowded, and how long a bed is kept occupied by a particular

patient has an impact on the waiting time of future patients, incurring system costs that go beyond what is experienced by the patient alone. Next, we expand on our single-patient/bed formulation in an effort to capture the system-level impact of decisions made for individual patients.

5.3. Incorporating ED Census into BeRT Decisions

The basic question we are interested in throughout this section is how the decision of whether to request a bed for a patient at the completion of triage should depend on the ED census at the time the decision is made. As we discussed before, our goal is to develop solution methods through the analysis of simplified formulations. Thus, to incorporate ED census into our decision framework, we first abstract away from the actual system, analyze a mathematical model that captures the very basic underlying dynamics, and establish a structure for “good” decision rules. Then, we develop a method for identifying and fully describing a policy that possesses such a structure. Later, in Section 5.5, we explain how we can fully operationalize this policy in practice.

5.3.1. Identifying the Structure of Good Policies: A Queueing Approach.

There are mainly three simplifications we bring to the problem. First, we assume that each ED bed has a separate stream of patients who line up for admission to that bed. Second, we assume that patients arrive according to a stationary process. Third, we assume that there is always a hospital bed that can be allocated to a bed request. None of these assumptions hold in reality. EDs have multiple beds, and waiting patients are admitted to these beds as they become available. Arrivals to the EDs are known to be highly nonstationary. Finally, there are finitely many hospital beds, and it is possible that all of these beds are full when an ED patient is admitted to the hospital. However, as long as the total patient load is allocated to each bed, it might be reasonable to expect that assuming a separate arrival stream may not be all that harmful. For the patient arrival process, given the difficulty of analyzing queueing models with nonstationary arrivals, one might proceed with assuming stationarity for the analysis and incorporate nonstationarity in a heuristic fashion postanalysis. Similarly, even though direct incorporation of hospital bed capacity will not be possible (because of analytical difficulties as well as our lack of access to the relevant data), as we explain later in the paper, dependence on hospital bed capacity can be captured at least partially through time-dependent estimation of TPP times.

Specifically, for our queueing analysis, we consider a single-server queue, where the server is meant to represent an ED bed. Similar to the model described in Section 5.1, we assume that patient arrivals follow a Poisson process with rate λ ($0 < \lambda < \infty$); triage times are negligible;

and thus, at the time of a patient's arrival, the hospital admission probability for the patient is known. Differing from the model of Section 5.1, however, patients are served according to the FCFS principle. There is also a single patient type. Therefore, we drop the type superscript we used in our earlier formulation and use Z_k to denote the random variable representing the probability that the k th patient to arrive will be admitted to the hospital. Just as before, we assume that $\{Z_k\}_{k=1}^\infty$ is a sequence of iid random variables with the common discrete probability distribution specified as $P\{Z_k = \alpha_i\} = q_i$ for $\alpha_i \in \Omega$ and $k \in \{1, 2, \dots\}$, where $\Omega = \{\alpha_1, \alpha_2, \dots, \alpha_M\}$ is the finite set of possible values Z_k can take. Without loss of generality, we assume that α_i is increasing in i . We also let $\alpha = E[Z_k] = \sum_{i=1}^M q_i \alpha_i$ so that α represents the probability that a randomly chosen patient, pretriage, will be admitted to the hospital.

The assumptions regarding workup and TPP times and costs are the same as in the single-patient model of Section 5.2. Specifically, letting S_k denote the ED workup time and B_k denote the TPP time for patient k , we assume that $\{S_k\}_{k=1}^\infty$ is a sequence of iid random variables with exponential distribution with rate γ_S and that $\{B_k\}_{k=1}^\infty$ is a sequence of iid random variables with exponential distribution with rate γ_B .

The decision to be made is that every time the server picks up a new patient, given the probability of admission for the patient and the number of patients in the queue, whether to request a hospital bed for the patient. If a bed is requested for patient k , whose probability of admission is z , it costs c_{tr} and the length of time the server (i.e., the ED bed) remains occupied by the patient would be $\max(S_k, B_k)$ with probability z and S_k with probability $1 - z$. If no bed is requested for the patient, then with probability z , a decision to admit the patient will still be made; because this decision will have been made after the workup is over, it will cost c_{ad} , and the length of time the server will be occupied by the patient will be $S_k + B_k$. With probability $1 - z$, no bed will be requested costing nothing, and the server will be occupied for S_k units of time. (Mostly for analytical convenience, we allow the decision maker to reverse their decision regarding the early bed request as the number of patients in the queue changes as long as the patient's workup in the ED is still going on and if already requested, the hospital bed for the patient has not already been prepared.) The objective of the decision maker is to minimize the long-run average cost in this queueing system.

We model this problem as an MDP. The state space \mathbb{X} can be described as $\mathbb{X} = \{0\} \cup \{(m, n) \mid m \in \Omega \cup \{r, p\}, n \in \mathbb{Z}^+\}$, where state 0 is the state where the system is empty, states (m, n) are the states in which there are $n \geq 1$ patients in the system (including the patient with the server), and m is specified as follows. In any state for which

$m = \alpha_i \in \Omega$ the workup for the patient with the server is either in progress or about to start, the patient's probability of admission to the hospital is α_i and no hospital bed is readily waiting for the patient to be transferred to. In any state for which $m = r$ the workup of the patient with the server is in progress and the hospital bed for the patient is ready (i.e., the TPP for the patient is complete). Finally, in any state for which $m = p$ the patient with the server is already done with her ED workup and her TPP is in progress. The action space is $\mathcal{A} = \{0, 1\}$, where zero corresponds to the decision of not requesting and one corresponds to the decision of requesting a hospital bed early with the restriction that no action is available in state 0 and action 1 is only available in states $x = (\alpha_i, n)$, where $n \geq 1$ for some $i \geq 1$ (i.e., when there is at least one patient in the system and neither the patient is already done with the ED workup and is waiting for the hospital bed to be available nor the requested hospital bed is ready and is waiting for the patient's ED workup to be over). A stationary policy is then defined as a mapping from the state space \mathbb{X} to the action space \mathcal{A} . In the following, we restrict ourselves to the class of nonidling policies, which imply that the ED bed is never kept empty (the server is never idle) as long as there are patients waiting.

Using uniformization, the continuous-time MDP formulation can equivalently be written as a discrete-time MDP. Let $\beta = \lambda + \gamma_S + \gamma_B$ denote the uniformization constant. We set $\beta = 1$ without loss of generality. For any $x \in \mathbb{X}$, let $h(x)$ denote the relative value or bias for state x . For expositional convenience, we further define $h(\alpha_j, 0) = h(0)$ for $j = 1, 2, \dots$, although $(\alpha_j, 0)$ is not an element of the state space \mathbb{X} . Finally, let g denote the long-run average cost under an optimal policy. Then, the optimality equations can be written as follows:

$$g + h(0) = \lambda \sum_{j=1}^M q_j h(\alpha_j, 1) + (\gamma_S + \gamma_B)h(0); \quad (5.2)$$

for all $n \geq 1$ and $\alpha_i \in \Omega$,

$$g + h(\alpha_i, n) = nc_w + \lambda h(\alpha_i, n + 1) + (1 - \alpha_i)\gamma_S \sum_{j=1}^M q_j h(\alpha_j, n - 1) + \alpha_i \gamma_S h(p, n) + \gamma_B \min \left\{ h(\alpha_i, n) + \frac{\alpha_i \gamma_S}{\gamma_B} c_{ad}, h(r, n) + \frac{\gamma_S + \gamma_B}{\gamma_B} c_{tr} \right\}, \quad (5.3)$$

where $h(\alpha_j, 0) = h(0) = \sum_{k=1}^M q_k h(\alpha_k, 0)$ for all j . For all $n \geq 1$,

$$g + h(r, n) = nc_w + \lambda h(r, n + 1) + \gamma_S \sum_{j=1}^M q_j h(\alpha_j, n - 1) + \gamma_B h(r, n), \quad (5.4)$$

$$g + h(p, n) = nc_w + \lambda h(p, n + 1) + \gamma_B \sum_{j=1}^M q_j h(\alpha_j, n - 1) + \gamma_S h(p, n). \quad (5.5)$$

We first establish the existence of the solution to the optimality equations and consequently, the existence of a stationary optimal policy. (The proofs of all the analytical results are given in the online appendix.)

Theorem 1. *Suppose that $\lambda\left(\frac{1}{\gamma_S} + \frac{\alpha}{\gamma_B}\right) < 1$. Then, there exists a finite constant g and a finite function $h(\cdot)$ that satisfy the average cost optimality equations as stated in (5.2)–(5.5). Furthermore, if π^* is a stationary policy that returns the action that minimizes the right-hand side of the optimality equations given system state x , then π^* is average cost optimal with average cost g .*

Theorem 1 essentially states that if $\lambda\left(\frac{1}{\gamma_S} + \frac{\alpha}{\gamma_B}\right) < 1$ (i.e., if the queue is stable under the policy of never requesting beds early), then there exists a solution to the optimality equations, and an optimal policy can be determined using this solution. Under this condition, we can further prove that there exists an optimal policy that has a threshold structure as outlined in the following theorem.

Theorem 2. *Suppose that $\lambda\left(\frac{1}{\gamma_S} + \frac{\alpha}{\gamma_B}\right) < 1$. Then, there exists an integer $N(\alpha_i)$ for each $\alpha_i \in \Omega$ such that the policy when the system is in state (α_i, n) requests a hospital bed for the patient if and only if $n \geq N(\alpha_i)$ is optimal. Furthermore,*

$$N(\alpha_i) := \inf \left\{ n \geq 1 : h(\alpha_i, n) - h(r, n) > \frac{\gamma_S + \gamma_B}{\gamma_B} c_{tr} - \frac{\alpha_i \gamma_S}{\gamma_B} c_{ad} \right\}.$$

The fact that there exists an optimal threshold-type policy as stated by Theorem 2 is not surprising. Nevertheless, the result formally confirms our intuition and provides a rigorous support for seeking policies of such structure. In fact, we can push this analytical characterization one step further and show that this threshold on the number of patients is weakly decreasing in the admission probability.

Theorem 3. *Suppose that $\lambda\left(\frac{1}{\gamma_S} + \frac{\alpha}{\gamma_B}\right) < 1$. Then, the optimal threshold $N(\alpha_i)$ is a nonincreasing function of $\alpha_i \in \Omega$.*

With Theorem 3, we know that the higher the hospital admission probability for a patient, the lower the bar (in terms of the number of patients in the system) for making an early bed request. The result also immediately implies that we can equivalently define the threshold on the admission probability rather than the number of patients. In other words, rather than aiming to find the optimal threshold level on the number of patients given the hospital admission probability, we can instead aim to find the optimal threshold level on the admission probability given the number of patients in the system.

As a result of our queueing analysis presented here, we can conclude that there is some support for a policy that sets a certain threshold level on the admission probability for making early bed requests and decreases this

threshold level as the number of patients in the ED increases. Such a policy would also intuitively make sense because one would expect quick turnover of ED beds to be more beneficial when ED is crowded. The question, however, is how one can determine what specific policy to use (i.e., how precisely the threshold levels should be determined) in practice. Even in the highly specialized queueing setting we considered here, our results do not provide a complete characterization of an optimal policy. Theorem 2 gives an expression for $N(\alpha_i)$, the threshold as a function of the admission probability α_i , but this expression is in terms of the bias function $h(\cdot)$, which is not known. An ideal solution would be to have a closed-form expression for the optimal threshold $N(\alpha_i)$ in terms of parameters, which can be estimated from the existing data. This does not appear to be doable. However, we were able to develop a heuristic solution, which is described by a closed-form expression for a threshold function and has a performance that is very close to the performance under the optimal threshold levels. The basic idea behind the heuristic rests on first solving the clearing version of the queueing model presented and then, adjusting the solution to account for the fact that the clearing model ignores future arrivals.

5.3.2. Clearing Version of the Queueing Model and its Analysis.

We introduce the following simplification to the model we studied in Section 5.3.1; there are $l < \infty$ patients initially in the system, and there will be no future arrivals. All other model assumptions remain the same, but in this case, we seek an optimal policy, which minimizes the expected total cost that will accumulate until the system is cleared of all the patients.

The server serves the patients in a random order. The admission probability for a patient is known only after the patient is chosen for service, and the server has to complete the service of a patient he or she has already chosen before moving on to another patient. The state space \mathbb{Y} can then be described as $\mathbb{Y} = \{0\} \cup \{(m, n) \mid m \in \Omega \cup \{r, p\}, 1 \leq n \leq l\}$, where m and n are defined as in the queueing model described in Section 5.3.1. Letting $V(y)$ for $y \in \mathbb{Y}$ denote the total expected cost that will accumulate until all the patients leave the system starting from state y , we can write the optimality equations as follows.

$$\begin{aligned} & \text{For } 1 \leq n \leq l, \\ V(\alpha_i, n) &= \frac{nc_w}{\gamma_S + \gamma_B} + \frac{(1 - \alpha_i)\gamma_S}{\gamma_S + \gamma_B} \sum_{j=1}^M q_j V(\alpha_j, n - 1) \\ & \quad + \frac{\alpha_i \gamma_S}{\gamma_S + \gamma_B} V(p, n) \\ & \quad + \frac{\gamma_B}{\gamma_S + \gamma_B} \min \left\{ V(\alpha_i, n) + \frac{\alpha_i \gamma_S}{\gamma_B} c_{ad}, V(r, n) \right. \\ & \quad \quad \left. + \frac{\gamma_S + \gamma_B}{\gamma_B} c_{tr} \right\}, \end{aligned}$$

where we let $V(\alpha_i, 0) = V(0) = 0$ for all $\alpha_i \in \Omega$ for notational convenience:

$$\begin{aligned} V(r, n) &= \frac{nc_w}{\gamma_S + \gamma_B} + \frac{\gamma_S}{\gamma_S + \gamma_B} \sum_{j=1}^M q_j V(\alpha_j, n-1) \\ &\quad + \frac{\gamma_B}{\gamma_S + \gamma_B} V(r, n), \\ V(p, n) &= \frac{nc_w}{\gamma_S + \gamma_B} + \frac{\gamma_B}{\gamma_S + \gamma_B} \sum_{j=1}^M q_j V(\alpha_j, n-1) \\ &\quad + \frac{\gamma_S}{\gamma_S + \gamma_B} V(p, n). \end{aligned}$$

The following theorem provides a complete characterization of the optimal policy.

Theorem 4. For any given state $y = (\alpha_i, n)$, it is optimal to request a hospital bed early if and only if

$$n \geq \frac{(c_{tr} - \alpha_i c_{ad})(\gamma_S + \gamma_B)}{\alpha_i c_w}. \quad (5.6)$$

The right-hand side of Inequality (5.6) provides a simple convenient formula for the optimal threshold on the number of patients, above which a hospital bed should be requested for the randomly selected patient. The threshold is a function of the patient's admission probability. However, it is determined under the assumption that no future patients will arrive. Next, we propose an approximate way of incorporating the reality that there will be future arrivals into the threshold formula given in (5.6).

The queue-length process of a single-server queue can be seen as consisting of a sequence of independent cycles, with each cycle defined by one busy period (during which there are customers in the system and the server is busy) and one idle period (during which there are no customers in the system and the server is idle). As it is clear from our queueing formulation, costs incur, and one can have some control on the costs only when the system is in the busy period. Therefore, it might be reasonable to believe that focusing on the minimization of the expected total cost that accumulates over busy periods might lead to policies that perform well under the long-run average cost minimization objective as well. Coming up with a precise description of the optimal policy that minimizes total costs over busy periods, however, is not any simpler than coming up with one for the queueing formulation in the first place. Nevertheless, it would be reasonable to believe (in part based on Theorems 2 and 3) that there must exist a threshold-type policy, which performs well if not optimally, and that the threshold should be lower when the admission probability for the patient is higher and when the arrival rate to the queue is higher (i.e., when the busy period is more likely to be long). As a heuristic solution, one can directly use the policy described in Theorem 4, but because the

clearing model ignores future arrivals, it underestimates the load on the system (more specifically, the number of patients who will receive service during the busy period), and as a result, the policy would most likely offer a threshold level that is higher than what it should be. To account for this deficiency, we propose that n on the left-hand side of (5.6) be replaced by $n/(1 - \lambda\tau)$, which is the expected number of patients who will be served until the first time the server is idle under the assumption that the expected time the server is occupied by each patient is τ . Note that this time in fact depends on the policy used, but for approximation purposes, we set $\tau = \left(\frac{1}{\gamma_S} + \frac{\alpha}{\gamma_B} - \frac{\alpha}{\gamma_S + \gamma_B}\right)^{-1}$, which is the expected time the server is occupied by a patient under the assumption that the hospital bed is requested at the time of triage. Thus, as a heuristic solution to the queueing problem, we propose the following threshold on the number of patients given the admission probability of a patient:

$$n \geq \frac{(c_{tr} - \alpha_i c_{ad})(\gamma_S + \gamma_B)(1 - \lambda\tau)}{\alpha_i c_w}, \quad (5.7)$$

or equivalently, the following threshold on the admission probability for the incoming patient given the number of patients currently in the system:

$$\alpha_i \geq \frac{c_{tr}}{c_{ad} + nc_w((\gamma_S + \gamma_B)(1 - \lambda\tau))^{-1}}. \quad (5.8)$$

It might be helpful to compare this threshold with (5.1), the threshold suggested by our single-patient/bed formulation $\left(z \geq \frac{c_{tr}}{c_{ad} + c_w(\gamma_S + \gamma_B)^{-1}}\right)$. Note that the difference between the two thresholds is in the multipliers for the waiting cost parameter c_w . The threshold for the queueing/clearing approach has the extra $n(1 - \lambda\tau)^{-1}$ multiplier capturing the effect of the existing number of patients in the system and the overall load.

Our numerical experiments, which are not provided here in the interest of space, showed that the performance of the policy that uses the threshold (5.7) (or (5.8)) within our queueing framework of Section 5.3.1 is very close to that of an optimal policy. Thus, our heuristic solution approach works very well within the confines of our queueing model. However, our ultimate objective is to develop a “good” solution for the actual system, not for the queueing model. To that end, in the next section, we describe how one can develop practical rules based on the results of this section and propose a specific policy.

5.4. A Heuristic for Deciding When to BeRT: CTT

There are several challenges in devising a “good” implementable policy based on our mathematical analysis. In particular, one needs to find a way to marry the stationary single-bed approach of our simplified mathematical models with the reality that EDs have multiple beds, they operate in highly nonstationary environments with

time-dependent patient arrival rates and workup times, and the timely availability of (or lack thereof) hospital beds might have a significant impact on boarding times. It is also not clear how one can set the values of the cost terms c_{wr} , c_{ad} , and c_{tr} in practice. We postpone the discussion of how to set the cost parameters until Section 5.5. Here, we explain how we get around the other challenges.

To capture the nonstationarity of patient arrivals to the ED, ED workup times, and boarding times, we propose using the time-based estimates of the model parameters as the policies are implemented in real time. For example, when using the threshold expressed in (5.1) in the policies we propose, if the decision is made at some time t where t corresponds to a specific one-hour window on a specific day of the week, in place of $1/\gamma_S$ and $1/\gamma_B$, we respectively use the estimates for the expected ED workup time and expected boarding time corresponding to that one-hour time window. (In fact, we push this one step further and allow the estimates to depend on not only time but also, the ESI and adult/pediatric classification of the patient.) This allows the policies to be more responsive to the changing dynamics throughout the week and through each day, and it helps capture the scheduled changes in staffing levels as well as hospital bed availabilities, which are known to vary significantly with time depending on the hospital's patient discharge practices (see, e.g., Shi et al. 2016). (Complete details on our estimates for the ED workup, boarding, and discharge times are provided in Section E.4 of the online appendix.)

Specifically, we divide each day of the week into 24 disjoint one-hour time slots, resulting in a total of $24 \times 7 = 168$ time intervals over each week, with the first time slot of each week corresponding to 12:00–12:59 a.m. on Sunday and the last slot corresponding to 11:00–11:59 p.m. on Saturday. Consider a patient who is admitted to her ED bed at time t , and suppose that her hospital admission probability (as estimated at triage) is z . Suppose also that at the time the patient is admitted to the ED bed, there are N patients in total, including the patient herself, and other patients in the ED occupying ED beds as well as those who are waiting either for triage or for an ED bed to become available posttriage. Let K denote the total bed capacity of the ED at time t . (The bed capacity and the staffing levels in most EDs, like in the ED where our data came from, change according to predetermined schedules.) Suppose that λ denotes the estimated new patient arrival rate for time period t . Let $1/\gamma_S$ and $1/\gamma_B$ denote the expected ED workup time and expected boarding time, respectively, corresponding to time period t , the patient's ESI level, and whether the patient is adult or pediatric patient. Thus, K and λ depend on t , whereas γ_S and γ_B depend on t as well as the ESI level and the age category of the patient, but we suppress this dependency in our notation for expositional simplicity.

One important issue to address is a potential unintended consequence of requesting hospital beds early when hospital capacity is extremely tight. If a hospital is too crowded for a long period of time to the extent that there is typically no hospital bed readily available when a bed is requested and almost all the hospital bed requests are put in a queue, then there is a possibility that because of the flooding of the queue with false early bed requests, the queue can be unstable, and as a result, boarding times of patients who are actually admitted can get increasingly longer. To prevent this from happening, it would be reasonable to turn off early bed requests whenever the hospital gets close to its full capacity, and therefore, as part of our policy, we propose that beds can be requested only when the number of hospital beds that are available for future ED patients is greater than or equal to some *predetermined protection level*, which conservatively, can be set to be equal to the expected number of hospital admissions on a given day.

Now, let us proceed with the development of the decision rules for our policy. Let M denote the number of hospital beds currently available for allocation to the ED patients who will be admitted to the hospital, and let A denote the predetermined protection level for the number of hospital beds. If $M < A$ at the time a patient goes through triage, then regardless of the hospital admission probability of the patient or the ED crowding level, a hospital bed will not be requested for the patient at triage. If $M \geq A$, then whether a hospital bed will be requested early for the patient depends on the patient's admission probability and the ED crowding level as explained in the following.

First, suppose that $N \leq K$ (i.e., there are no patients waiting for an ED bed to become available). In this case, it would be reasonable to assume that the impact of the early bed decision made for the patient on the future patients would be minimal, and thus, it would make sense to use a decision rule that focuses on a single patient alone as we did in Section 5.2. Thus, the policy we propose calls for using the threshold in (5.1).

Suppose now that $n > K$. In this case, there are already patients waiting for an ED bed to become available, and therefore, it will be important to capture the queueing dynamics so as to factor in the impact of decisions made for a particular patient on the waiting times of the future patients. It would be reasonable to assume that as long as $n > K$ because all the beds will be occupied (all "servers" are busy), the evolution of the queue-length process in this multiserver queue will be like the evolution of a single-server queue, with the service speed of the server being K times the speed of a single server. Therefore, it would be reasonable to use a policy that is based on the threshold (5.6) obtained in Section 5.3.2. One problem with approximating the multiserver queue with the single-server one as described is that the single-server queue overestimates the percentage of time there

are more than K patients in the ED (because in the single-server queue when $N < K$, there is in effect no one in the system, and consequently, there is no service), and as a result, the decision rule with this approximation would likely be more inclined to request beds early by setting a threshold level lower than it should. Therefore, it would be reasonable to make an adjustment on this threshold by shifting the threshold curve (as a function of N) upward so that when the function is evaluated at $n = K$, it will match with the single-patient threshold of (5.1). This way, the threshold, as a function of N , would be a continuous function with a flat portion for $N \leq K$ and a convex decreasing portion for $N > K$. With this adjustment in place, CTT can be described as follows.

5.4.1. Description of the CTT Policy.

If $M < A$, request a hospital bed for the patient, if needed, only after the disposition decision for the patient.

If $M \geq A$ and $N < K$, then request a hospital bed for the patient at the completion of triage if the patient is an ESI-1 or ESI-2 patient and

$$z \geq \frac{c_{tr}}{c_{ad} + c_w(\gamma_S + \gamma_B)^{-1}}; \quad (5.9)$$

otherwise, request the bed, if needed, after the disposition decision for the patient.

If $M \geq A$ and $N \geq K$, then request a hospital bed for the patient at the completion of triage if the patient is an ESI-1 or ESI-2 patient and

$$z \geq \frac{c_{tr}}{c_{ad} + (N - K)c_w(K(\gamma_S + \gamma_B))^{-1}(1 - \lambda\tau)^{-1}} - \frac{c_{tr}c_w}{c_{ad}(c_{ad}(\gamma_S + \gamma_B) + c_w)}; \quad (5.10)$$

otherwise, request the bed, if needed, after the disposition decision for the patient.

5.5. Setting the Parameters of CTT

There are no easy answers to the question of how one should set the cost parameters of CTT. As we explained at the beginning of Section 5, these cost parameters are not meant to be taken in dollar terms but rather, as weights that help capture the trade-off between the number of false bed requests and the average ED length of stay, the two main performance measures of interest. It does not seem to be difficult for practitioners to recognize that what matters is not the actual values of these parameters but rather, their values relative to each other, and these parameters can best be seen as tools for making adjustments and fine-tuning the policy once it is put in place. Nevertheless, one has to have at least some rough idea as to what values should be considered at the very least for starting the implementation and guiding any future adjustments.

There appears to be no discernible reason as to why c_{tr} and c_{ad} should be different from each other. It would also be reasonable to normalize their values by setting $c_{tr} = c_{ad} = 1$ and focus on how one should set c_w accordingly. Physicians and other key decision makers in the ED cannot directly come up with an estimate for c_w , but we can start with the assessments and opinions that they would be relatively comfortable with and that roughly capture the main trade-offs between the “cost” of making incorrect early bed requests and potential operational gains and use them to come up with an estimate for c_w . For example, we found out that ED physicians would be relatively comfortable with making an assessment as to how many true early bed requests would be needed on average for every false early bed request and used this assessment to come up with a range for c_w . When making this rough assessment, physicians need to weigh the potential benefits of making early bed requests with the potential negative reaction that false requests would get from the hospital staff and management, which could also potentially culminate in resistance to the adoption of the early bed request practice with time.

Let m denote the number of true early bed requests that we need to have for every false early bed request, and let D_{LOS} denote the random variable denoting the difference between the ED length of stay without early bed request and the ED length of stay with early bed request for a random patient. Then, if we ignore the impact of a true bed request on the patients in the ED other than the patient for whom the bed is requested, the expected total length of stay “savings” from m true early bed requests would be $mE[D_{LOS}]$. We can then argue that if m true early bed requests are equal in value to one false early bed request, we must have $c_w mE[D_{LOS}] = 1$, and thus, we can evaluate c_w in terms of m by

$$c_w = \frac{1}{mE[D_{LOS}]}.$$

Obviously, this formula will not give us an absolute value for c_w because apart from the approximations we made, there is not a definite answer as to what m should be. Furthermore, because we ignored the impact of the early bed request on the other patients, c_w , as computed, should be taken more as an estimate for a lower bound. Therefore, it would be reasonable to start with a commonly agreed upon choice for m , use the formula to get a lower bound for c_w , and then, consider a range of choices that are larger than c_w .

Using our data from the ED, we estimated that $E[D_{LOS}] = 1.98$ hours for a random patient. When making this estimation, we had to make an assumption as to how long TPP for a random patient would last when the request is put in early because there are no data that one can use to make this estimation. Specifically, we assumed that TPP for the patient would be the same as what the boarding time for the patient would

have been if the bed was not requested early. We also kept the distributional assumptions we made for the ED workup and boarding times in our mathematical model of Section 5. The choice of $m = 10$ was deemed to be reasonable by one of the coauthors of this paper who is an emergency medicine physician and the vice chair of strategic initiatives and operations at a hospital, which implied $c_w = 0.051$ units per hour. In our simulation study, we picked c_w values within a range that is slightly above 0.051, specifically within the range of 0.06–0.26. In Section 6, we provide more explanation for this choice.

5.6. FT

There are no significant obstacles to implementing CTT in practice. One can easily integrate it as a decision support tool with the existing electronic healthcare record system in place, and using the simple formulas provided in the description of CTT in Section 5.4, the tool could easily and quickly determine whether conditions (including the health condition of each patient as well as ED census levels) justify requesting a bed early and alert the appropriate ED staff to initiate TPP.

Nevertheless, it can still be potentially desirable to use even a simpler policy: for example, one that sticks with a single threshold level on the probability of admission 24 hours a day regardless of the changing ED census as long as hospital beds are not in short supply (i.e., as long as the number of hospital beds that are available for allocation to newly admitted ED patients is greater than the predetermined level of A). We call this policy the FT. (Note that just like ESIB and CTT, FT also considers ESI-1 and ESI-2 patients only for a potential early bed request.) The obvious question when implementing this policy is what exactly the fixed threshold should be. One possibility is to set the threshold at all times to the right-hand side of (5.9) (i.e., essentially assuming that each patient is independent of the others or that there are never patients waiting to be admitted to the ED) using time-independent overall estimates for γ_S and γ_B and setting the cost parameters as discussed in Section 5.5. However, it would be more reasonable to use the threshold obtained this way as a starting point for determining what threshold to use and make adjustments accordingly. In our simulation study, which we describe in the following section, we considered a range of values for the fixed threshold value so as to make a fair comparison among the two policies we propose and the benchmark policy ESIB.

6. Simulation Study

The real test of how much benefit requesting beds early at the time of triage would bring and which one of the two policies performs better would be through implementing these policies in practice. However, because implementation of ESIB, FT, or CTT would mean a major

change in the way ED-hospital operations are run, implementing these two policies for testing purposes was not an option. Therefore, we carried out a discrete-event simulation study instead. In this section, we report our findings.

Developing valid and useful simulation models for complex systems is a significant challenge. The main difficulty, as explained clearly in chapter 5 of Law (2007), is to find the right level of detail that should be captured in the model. Obviously, the model should be a relatively close representation of the actual system, but this does not mean that a model that uses a more detailed and supposedly more “realistic” approach is better. The right level of modeling depends on many factors, including the goal of the study and the available data elements that can be used to populate model parameters. As a result of a months-long process during which we followed the guidelines provided in Law (2007), we converged on a model, carefully calibrated its parameters, and carried out formal model validation. Complete details of the model as well as a summary of the calibration and validation process are provided in Section E of the online appendix. Here, we provide a broad description highlighting the important features of our simulation model, which relax many of the simplifying assumptions of our mathematical models. In particular, unlike the simplified mathematical models we used to devise our policies, in our simulation analysis, we model the finite hospital capacity explicitly. The description is for the ED operating under the current policy of no early bed requests, but as explained in the online appendix when simulating the ED with possible early bed requests, we use the same underlying model.

In the simulation model, patients of each ESI level arrive at the emergency department according to a non-homogeneous Poisson process with rates that depend on the day of week as well as the time of day. Each arriving patient goes through triage, at the end of which the patient’s ESI level is revealed and the probability of hospital admission is determined. The ED has a finite number of beds, and the number of beds available depends on the time of day. As long as there are beds available in the ED, patients are admitted to the ED right after triage. However, if there are no beds available, patients wait until one is available, with priorities determined according to patients’ ESI levels. (Complete details on how prioritization works and which patient is admitted to which pod in the ED are given in the online appendix.) A patient admitted to the ED goes through two stages of service. The first stage corresponds to ED workup, during which the ED personnel members carry out all the tasks needed to reach a diagnosis for the patient and perform any urgent treatment that can be done in the ED. This first stage concludes with a disposition decision that requires the patient either to be discharged from the ED or to be admitted to the hospital. If the decision is to

discharge the patient, then the second stage corresponds to the discharge process for the patient. If the decision is to admit the patient to the hospital, then the second stage corresponds to the patient's boarding time. Following an approach that is similar to that of Shi et al. (2016), we model the boarding time of a patient so that it depends on the availability of the hospital beds and the TPP time for the patient. Specifically, when the admit decision is made, if a hospital bed is available, then the boarding time simply equals the TPP time; however, if there are no beds available, then the hospital bed request joins a queue and waits for a bed to become available. (As patients are discharged from the hospital and beds become available, they are allocated to the requests in this queue in an FCFS manner.) In this case, the boarding time of a patient is equal to the waiting time in the queue plus the TPP time. Once the second-stage service is over, the patient leaves the ED, making an ED bed available for patients who are waiting or will arrive in the future. Patients who are transferred to the hospital keep a single bed occupied during their stay. See Section E.4 in the online appendix for details on the probability distributions for ED workup time, TPP time, and hospital length of stay.

We considered mainly two different scenarios for the simulation study. First, we assumed that the ED consistently operates under "normal" operating conditions as estimated using historical data (described in Section E in the online appendix), and we used long-run average analysis for making performance comparisons (Section 6.1). Then, we considered an alternative setting where we assumed that over a period of six weeks, the ED experiences unexpectedly high patient volumes, possibly because of an outbreak, and we compared the performances of the alternative policies over a longer period that includes that six-week period (Section 6.2).

One important issue is deciding what criterion or criteria to use when comparing policies or investigating whether they would improve upon the current practice. As we explained at the beginning of Section 5, despite its unappealing vagueness, a more easily understood goal in practice would be something like "reduce patient length of stay but do not have too many false bed requests." It is difficult to work with such an evaluation criteria, but in the following, we make an effort to try to get to such an evaluation as much as possible. More specifically, for performance comparison, we mainly consider average length of stay and average number of false bed requests per day (although we also report and comment on other performance measures, such as average length of stay for admitted patients alone and total number of early bed requests per day). One possible way to take both of these criteria into account is agreeing on a fixed value for the number of daily false bed requests and then, comparing the policies with respect to the average length of stay alone. However, not only is it difficult to

configure CTT and FT so that they will hit the desired level of false bed requests precisely, what level of daily false bed requests would be considered reasonably small is not clear in the first place. This level would likely change depending on the ED as well as the conditions it is operating in. Therefore, rather than setting a specific level for the acceptable value of the average number of daily false bed requests, we consider a range of possibilities and compare the policies by identifying the efficient frontier and observing whether this frontier is dominated by any particular policy.

6.1. Scenario 1: Long-Run Average Analysis Based on Historical Estimates

For this study, we used a simulation model precisely as described in Sections E.1 and E.3 of the online appendix and made comparisons over long-run horizon averages. Currently, no early bed request system is in place, and therefore, the average number of false bed requests per day is zero. The average ED length of stay per patient is approximately 356 minutes, and the average ED length of stay for patients who are admitted to the hospital is approximately 532 minutes (based on data collected during 2012). If we use the policy of requesting beds early for all the patients regardless of their ESI levels, a policy that cannot possibly be implemented, we find from our simulation analysis that the average length of stay would be approximately 318 minutes, with a 95% half-width of 0.6 minutes (the average for admitted patients would be 441 minutes), and the expected number of false bed requests per day would be 66. It is useful to know this number (318) in minutes to get an idea about how small the average length of stay could potentially be, but given that getting to that figure would require 66 false requests per day, it is safe to say that in reality, it would be impossible to get anywhere close to it. A reasonable range for the number of false requests per day, at least for an ED that is similar to the one we consider here with an average total patient load of roughly 180 per day and an average number of daily hospital admissions between 30 and 40 depending on the day of week, would be zero to three (at most up to five under exceptional circumstances). The question then would be, within this range, how much improvement one would get in terms of average length of stay by using early bed request policies and which specific policy would lead to the smallest average length of stay. To answer these questions, we simulate the system under FT using a range of values for the threshold ξ and simulate the system under CTT using a range of values for c_w . For ESIB, we have not specified any policy parameters. However, one can obtain a range of performances for ESIB by choosing to turn it on and off during different time periods of the day rather than keeping it on the whole day. Thus, to be able to compare FT and CTT with ESIB, we simulate the system under ESIB under different settings each characterized by the

time interval during which ESIB is kept on. For each one of the three policies, we compute the values for the two performance criteria and determine whether the efficient frontier is dominated by any one of the policies. Figure 1 is a visual demonstration of how the three policies compare.

When constructing Figure 1, we conducted seven separate simulation experiments under FT with seven different choices for the threshold ξ and seven simulation experiments under CTT with seven different choices for c_w . The choices for ξ were 0.91, 0.84, 0.81, 0.785, 0.736, 0.71, and 0.692, so that the expected number of daily false bed requests covered the range from zero to three. The choices for c_w (in terms of units per hour) were 0.06, 0.11, 0.14, 0.16, 0.2, 0.24, and 0.26. Recall from Section 5.5 that our rough estimate for a lower bound on c_w was 0.051, and thus, our choices for c_w cover a range of plausible values for c_w , with 0.051 taken as a lower bound. As to why we chose these seven specific values for ξ and c_w , we had the following goal; we aimed to have them somewhat evenly spaced over the respective intervals, but we wanted to do this in a way that the expected number of daily false bed requests under FT for a given threshold value matched with the expected number under CTT for a given value of c_w so that we could compare the two policies over their performances with respect to the average length of stay. For ESIB, we considered four different settings by considering four different time intervals during which the policy would be implemented. Specifically, we used the intervals 3–4, 2–4, 1–5, and 12–6 p.m. Note that these intervals are chosen within the busiest time periods of the day as we have found that ESIB makes the most difference when implemented when the ED is busy.

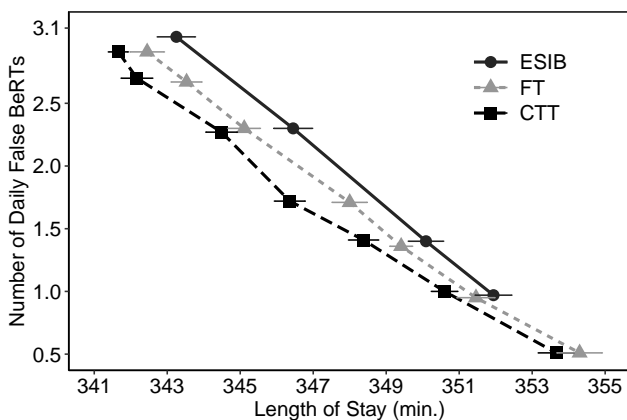
In Figure 1, the marks (circles for ESIB, triangles for FT, and squares for CTT) represent the mean number of daily false bed requests and the mean value for the

average length of stay, and the error bars around each point represent the 95% confidence intervals for the average length of stay. Because the confidence intervals for the expected number of daily false bed requests were very small, we did not show them on the figure. When simulating the system under all three policies and different policy parameter choices, the warm-up period was 366 days (based on a warm-up period analysis), and we used a batch means method with a total simulation length of 365,250 days.

From Figure 1, we can observe that even with an average number of daily false requests of one, there could be an approximately five-minute reduction in the average length of stay per patient over the status quo. Note that this average is over all the patients who arrive at the ED, not just patients who are admitted to the main hospital. If we look at the reduction in average length of stay for admitted patients only, the reduction is approximately 10 minutes. (See Figure EC.3 in Section E.5 of the online appendix.) At first glance, a five-minute reduction may not seem important. However, considering that on average more than 180 patients visit the ED on a given day, a five-minute improvement on average would be significant. This would roughly correspond to having an ability to see two to three more patients every day at the current (year 2012) levels of patient length of stay. If the ED is willing to go up to three incorrect bed requests per day, then the decrease in the average length of stay would get close to 10 minutes, corresponding to an additional capacity of about five patients per day. In short, our study suggests that requesting beds early at triage has the potential to make significant improvements in patient length of stay.

One important question is whether the two policies we propose, FT and CTT, make a difference over the benchmark policy ESIB. We can see from Figure 1 that they do. To be clear, even though the mean performance of FT appears to be better than ESIB in all the scenarios considered, FT is not always statistically superior to ESIB (at the 0.05 level of significance). However, CTT appears to be statistically better than ESIB across the board. Recall that both FT and CTT restrict the early bed request decision to ESI-1 and ESI-2 patients only. One question of interest is how their performances would be impacted if this restriction is lifted and patients from any ESI level would be eligible under FT and CTT. Figure EC.2 in Section E.5 in the online appendix answers that question. As we can see from the figure, which is constructed using different ξ and c_w values when using FT and CTT for all ESI levels (to achieve daily numbers of false BeRTs that are comparable with those under the other policies), when the two policies are restricted to ESI-1 and ESI-2 patients, their performances are statistically better. In fact, we tested using FT and CTT under different sets of restrictions (making ESI levels 1–3 eligible, 1–4 eligible, etc.), and we found that the best performance is clearly

Figure 1. Long-Run Average Length of Stay and Long-Run Average Number of Daily False Bed Requests Under ESIB, FT, and CTT for Scenario 1



achieved when the policies are restricted to ESI-1 and ESI-2 patients only. (These plots are omitted for brevity.)

If we compare the performances of CTT and FTT, we can see from Figure 1 that CTT appears to either dominate or be equivalent to FT over the range of practically feasible policy parameter settings considered. In most cases, the performances of CTT and FT appear to be statistically indistinguishable. However, it is notable that in some cases, CTT is statistically superior, and its mean performance is consistently better than that of FT. If we look at how big of an improvement CTT brings over FT, we see that the improvement in the mean length of stay can be up to two minutes. Even such an improvement would be important, especially because CTT essentially comes with no extra “cost.” It is no more difficult to implement than FT because both policies would require adding a simple decision support tool to the existing electronic medical record system in place. It is also not clear whether the supposed potential advantage of FT over CTT, which is that the admission threshold is fixed at all times, is all that important. This is because whatever threshold level the decision support tool is using at any given time will be largely invisible to the staff, and thus, there does not appear to be any clear reason why changing the threshold level based on system conditions would have some adverse effects. Still, there could be some hidden benefits to using a simpler policy, like FT, and thus, one might question whether it is worth skipping it in preference to CTT.

Note that in Section E.5 of the online appendix, we also provide plots of both the overall ED patient length of stay and the ED length of stay for admitted patients only with respect to the average number of daily total early bed requests (not just false requests). We can see that plots for the average daily total early bed requests are similar to those we have for the false early bed requests, and thus, one can reach the same conclusions, even if the focus is on keeping the total early requests low as opposed to only the false ones.

There is, however, one important issue to consider before reaching a final determination as to whether one policy appears to be better than the other. For the simulation scenario we considered in this section, we assumed that every week, the ED experienced a stochastically identical patient arrival process. As explained in Section E.1 of the online appendix, this arrival process is found as a good fit based on the historical data and thus, generally captures the reality of the ED where our data came from, but nevertheless, this fit is still on average. Through statistical estimation, we essentially constructed a typical week for the arrivals to the ED and assumed that this typical week repeated itself over and over again. (To be clear, the arrival process is stochastic, and thus, the realizations of the arrival process are not identical across weeks; however, they are identical stochastically.) In reality, however, every now and then there are shifts in the patient demand the ED

observes, which would make a typical weekly arrival process a poor fit to the actual arrival process observed during or after the shift. For example, with the flu season, EDs typically observe an increase in patient arrivals, and this causes them to live with elevated demand levels for a period of time. It is thus also of interest to investigate the performances of CTT and FT over such a period when the ED is hit unexpectedly by a more than usual level of patient demand. This is what we consider in scenario 2.

6.2. Scenario 2: Finite-Horizon Analysis with an Elevated Arrival Pattern

The Centers for Disease Control and Prevention has developed a tool called FluSurge 2.0 (available from its website (Centers for Disease Control and Prevention 2017)), which can be used to predict the increase in demand hospitals will observe in the case of an influenza pandemic. The tool provides the user several options for creating different pandemic scenarios. The possible choices for the duration of the pandemic are 6, 8, and 12 weeks. During the first half of the pandemic, every day the arrival rate of patients increases by some a percentage compared with the previous day, peaks right in the middle of the pandemic’s total duration, and then, decreases by a percentage every day compared with the previous day during the second half of the pandemic. The default value for a is three. Note that our goal in this section is to consider a plausible scenario under which the ED experiences a more than usual level of patient demand for one reason or another over a period of a few weeks, not necessarily to consider pandemic-level conditions. Nevertheless, we can take the assumptions of FluSurge 2.0 as a starting point and modify them somewhat to construct alternative scenarios under which patient demand is elevated, even though this elevation may not be as assumed by FluSurge 2.0.

Specifically, for scenario 2, we simulate the ED over a period of 18 weeks only. (The warm-up period for each run was 60 days, whereas the number of replications was 500.) During the first and last six weeks, the arrival process is exactly the same as that assumed for scenario 1. The middle six-week period is what we call the *outbreak period*. During the first three weeks of the outbreak period, the arrival rates for each patient class corresponding to each day of week increase by 1.5% every day, and during the last three weeks of the outbreak period, the arrival rates for each patient class decreases by 1.5% every day. Thus, the arrival rates peak in the middle of the outbreak period and go back to the first six-week period levels by the time the last six-week period starts. Note that the day of the week still affects the arrival rate because in the simulation study, we use the arrival rates corresponding to whichever day of the week the simulation is in. For example, if the n th day of the outbreak period, with $n < 21$, is a Monday, then the arrival rates for that day are $(1.015)^n$ times of whatever the arrival rates are for a typical

Monday (i.e., Mondays outside the outbreak period). In this scenario, we consider two different cases for the main hospital’s surge capacity, with the hospital’s bed capacity increased to 400 or 450 (from the baseline cases of 310 and 350 beds depending on the time of day as assumed in scenario 1), whereas the ED’s bed capacity remains the same. (Hospitals do resort to such capacity increases in cases of emergencies that overwhelm their regular capacities. We should note that without the surge capacity, our policies would not make a difference.) Figure 2 shows the performances of ESIB, FT, and CTT over the 18-week period the system was simulated under the four different assumptions for the hospital bed capacity.

Because we do not have ED data on patient length of stay for the outbreak scenario considered here, we cannot make strong claims on what the average length of stay would be under the current policy of not making any early bed requests. However, in our simulation study, we found that when the inpatient bed capacity was 400 and no beds were requested early at triage, the mean value for the average length of stay was 543 minutes, with a 95% half-width of roughly 10 minutes. (When inpatient bed capacity was 450, the average length of stay and the approximate 95% half-width were, respectively, 514 and 8 minutes.) Thus, we can see from Figure 2 that all three policies would bring substantial improvements even if the ED is comfortable, with only one false bed request per day on average. Furthermore, unlike the case in scenario 1, for higher levels of acceptable false early bed requests at least up to a level of five false bed requests per day, the average length of stay continues to decline substantially, particularly under CTT. For example, when the hospital bed capacity is 400, the average length of stay is around 500 minutes when the average number of daily false bed requests is close to one under CTT; if the ED is willing to go up to three false requests per day on average, then the average length of stay drops to less than 460 minutes.

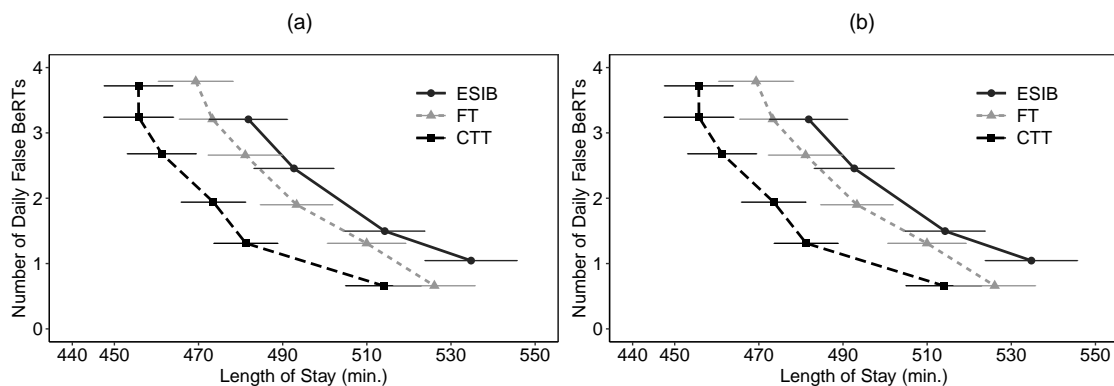
If we compare the performance of the benchmark policy ESIB with those of FT and CTT, we can see that both FT and CTT outperform ESIB. The performance of CTT is statistically and substantially better than the performance of ESIB across the board. As for FT, its performance is also statistically better than ESIB when the number of hospital beds equals 450. When the number of hospital beds equals 400, even though the performance of FT is not statistically better than that of ESIB, its mean performance is still consistently better than ESIB. Overall, we can conclude that the policies we propose bring substantial benefits over the benchmark policy, particularly when ED experiences higher than usual patient load.

Finally, we can also observe from the figure that there is a very clear difference between the performances of CTT and FT. It is not only that the performance of CTT is statistically better than that of FT, but also, the difference in their means is substantial. For example, with an average number of false bed requests of two, the difference between the average length of stay under CTT and the average length of stay under FT would be roughly 25 minutes with an inpatient bed capacity of 400, whereas it would be close to 30 minutes with an inpatient bed capacity of 450. In short, it appears that our state-dependent policy CTT appears to be much more responsive to unexpected fluctuations in crowding levels, which might be a result of unexpected deviations from the regular patient arrival patterns.

In Section E.6 of the online appendix, we also provide plots for both the overall ED length of stay and the ED length of stay for admitted patients only with respect to the total number of daily early bed requests. As in the case of scenario 1, one can make observations that align with what we noted based on plots for the daily number of false bed requests.

Going over the simulation results for both scenario 1 and scenario 2, we can identify three main takeaways. First, it appears that requesting beds early at triage for patients who have a high probability of being admitted

Figure 2. Average Length of Stay Vs. Average Number of Daily False Bed Requests Under FT and CTT for Scenario 2



Notes. (a) Number of beds = 400. (b) Number of beds = 450.

to the hospital has the potential to significantly improve patients' length of stay, and the improvements would likely be more substantial over periods when the ED observes elevated levels of patient arrivals. Second, the two policies we propose, FT and CTT, appear to bring significant improvements over the benchmark policy ESIB, again particularly when the ED experiences a high rate of patient arrivals. Third, CTT, by taking into account dynamically changing ED census levels, appears to be much more responsive to changing external conditions that impact patient demand, leading to substantially lower levels of patient length of stay for fixed levels of daily average false bed requests when compared with both FT and ESIB.

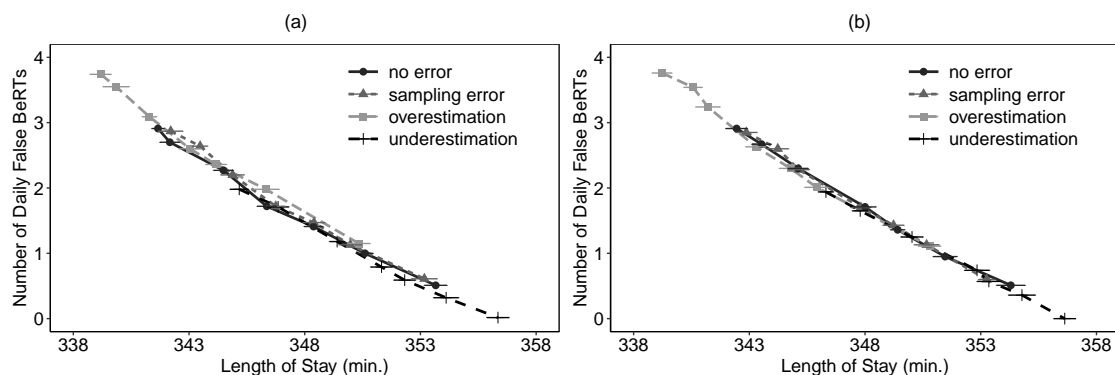
6.3. Scenario 1 Revisited: Errors in Admission Probability Estimates

One implicit assumption we made in our analysis so far was that the admission probabilities were correctly estimated. In reality, however, errors in estimation are inevitable, and thus, it is of interest to investigate their impact on the performances of our policies. In this section, we are interested in how the performances of our policies would change if there is *sampling* or *systematic* error in the estimation of admission probabilities. We experimented with three different settings. In one setting, we assumed that there was only sampling error, and thus, the actual admission probability for a patient was sampled from the truncated normal distribution with a standard deviation equal to 10% of the mean. In the remaining two cases, we assumed that there was a systematic error with consistent overestimation for all patients in one case and with consistent underestimation for all patients in the other. More specifically, in the overestimation case, the true admission probability was 90% of the estimated admission probability, and in the underestimation case, the true admission probability was 110% of the estimated admission probability truncated at one.

Figure 3 shows how the performances of CTT and FT are impacted by the admission probability estimation errors. To aid in comparison, we also plotted the performances for the case without any errors. (The no-error cases are the same as those plotted in Figure 1.) We can see from the figure that there is no distinguishable difference between the performances when there is no error versus when there is sampling error. When the admission probabilities are systematically overestimated, we can see that the performance curve is shifted to the left and upward compared with the no-error case, meaning that the policy would lead to more incorrect bed requests and consequently, lower ED length of stay. On the other hand, when the admission probabilities are systematically underestimated, the performance curve is shifted to the right and downward, meaning that the policy would lead to fewer incorrect bed requests and consequently, higher ED length of stay. This is not surprising because when admission probabilities are overestimated, the admission probability thresholds are exceeded by some of the patients who normally would not, and as a result, beds are requested early for more undeserving patients, leading to an increased number of daily incorrect early bed requests but also, an increased number of correct early bed requests, which leads to improvement in ED length of stay. When admission probabilities are systematically underestimated, some of the patients who would normally qualify for early bed requests do not, and as a result, fewer early bed requests are made, leading to lower errors but also, higher lengths of stay because underestimation also causes some of the true admits to be missed.

This analysis shows that our proposed policies will likely not be impacted significantly by sampling errors that are not too large. (Obviously, if the estimation can be significantly off having high variance, our policies would not perform well, just as any policy would not when implemented with incorrectly estimated policy parameters.) On the other hand, if the estimation error is

Figure 3. Average Length of Stay and Average Number of Daily False Bed Requests for Scenario 1 with Errors in Admission Probabilities



Notes. (a) CTT. (b) FT.

systematic, then the performances will be significantly impacted. However, in that case, the error should be obvious to the ED managers relatively quickly, and corrective action can be taken. As we explain in Section 5.5, even when estimation probabilities do not have errors, the policy parameters (the cost terms in the case of CTT and the threshold in the case of FT) are not fixed, predetermined values. In implementation, proper adjustments would need to be made so that the ED will operate in a way that works well in practice, carefully balancing the reduction in length of stay with incorrect bed requests. Therefore, in practice, if there is a systematic estimation error, the ED managers would make adjustments either in their probability estimates or directly in the policy parameters so that the ED will operate in the regime they find to be ideal. In other words, in the case of systematic errors, the underestimation and overestimation curves in the figures can practically be adjusted with time so that the impact of estimation errors would be negligibly small.

7. Concluding Remarks

The importance of improving patient flow in emergency departments and hospitals and consequently, reducing patient wait times is clear. There also appears to be a wide consensus on the idea that meaningful improvements in ED waiting times and lengths of stay can only be achieved through a systems approach that views the ED and the main hospital together as opposed to two distinct units. This essentially necessitates rethinking some of the traditional ways EDs and hospitals operate and developing new policies and procedures that require coordination between them. However, to the best of our knowledge, concrete ideas about how such new policies and procedures would look and scientifically rigorous studies on their potential impact have been largely missing. This paper makes a contribution toward filling this gap.

The paper is built on a simple and intuitively sound proposition; for patients who are likely to be admitted to the hospital, it might be worth it to plan ahead and prepare their hospital beds as the patients go through their examination and treatment in the ED. It is, however, not quite clear whether a patient's estimated probability of being admitted is high enough to justify requesting a bed early and whether and how the answer to this question should change with changing system conditions. We provided answers to these questions in this paper. In particular, we came up with specific policy prescriptions that clearly outlined how to decide whether a hospital bed should be requested for a patient at the time of triage. We found that the improvements in average patient length of stay with early bed requests could be significant, and the policies we propose have the potential to perform significantly better than a simple benchmark policy, which can readily be implemented in any ED that uses ESI classifications for triage with no further need for

a policy development. We also found that policies that make early bed request determination for a patient based on patient-specific information as well as changing census levels have the potential to work significantly better than policies that only use patient-specific information.

The policies we develop, even the state-dependent CTT, are easy to implement. Even though we used a specific ED as our data source, these policies can easily be used in any other ED because the general structure we assumed for the patient flow is shared by many (if not all) EDs. The only additional work needed would be to estimate key model parameters, such as arrival rates, by analyzing historical data. Integration of our methods with the existing electronic healthcare record systems is also not a big challenge. One potential issue could be that when it comes to implementation, for large academic hospitals with many admitting services, it might be necessary to estimate not only the probability that patients will be admitted to the hospital but more specifically, to which service and/or unit of the hospital they will be admitted. Even if that is the case, however, it is not difficult to develop another regression model to make that estimation. In short, there do not appear to be serious technical obstacles toward making "good" early bed request decisions, in particular using the policies we developed in this paper.

There is one major obstacle, however, and that is largely a cultural one. The idea of requesting beds early at triage seems simple but would in fact be a major change if implemented. The ED and the hospital staff are deeply accustomed to years of practice of starting the hospital admission process only after the patients' ED evaluation and treatment are over, the results of any blood or other diagnostic tests are available, and nothing else is left to be done for the patients in the ED. Therefore, it would be reasonable to expect that the immediate reaction, particularly from the main hospital staff, will be one of resistance to such a change. There will be some unease with the idea of making preparations for a patient for whom a complete picture is not known and who may in fact eventually not be admitted to the hospital. Even if the policy is implemented, early on, particularly if there are many false early bed requests, the staff involved might lose faith in the new policy, possibly resulting in long patient transfer preparation times and ultimately, resulting in the new policy's failure. Therefore, for success, prior to any implementation, buy-in from all the major stakeholders is essential. First, the hospital management should be convinced that the newly proposed policy has significant potential to make positive change. Second, with the management's help, the ED and the hospital staff should be educated in what way the new policy will make a difference, what the challenges will be, and what would be reasonable to expect when moved to implementation. Needless to add

of course, when moved to implementation, the proposed policies should start delivering what they promised quickly as otherwise, there will be significant push toward going back to the business as usual, possibly making it even more difficult to test such policies in the future. In short, two things are essential before any implementation: a policy whose superiority we are absolutely confident of and a systems-level trust in the policy, which can only be won through educating the staff and a scientifically sound, simulation-based demonstration of the policy's expected performance.

The change we demand from the practitioners is significant. Therefore, a single paper will not and should not be sufficient to make a strong-enough case that the hospitals should move toward implementation tomorrow. However, the potential for improvement is strong, and the policies we propose, even if they are not implemented exactly as we describe in this paper, can help inspire new ideas and form the basis for alternative policies in the future. It is worth noting that we do not make any claims in regard to the optimality or near optimality of the policies we propose, and we have no doubt that future work will improve upon the policies we propose. One interesting idea would be to consider making early bed requests not only at the time of triage but over the course of the patients' sojourn in the ED until their disposition decisions. Any uncertainty present at the time of triage regarding a patient's eventual disposition might diminish during the patient's workup, and so, it might be preferable to wait some time before deciding whether a hospital bed should be requested for the patient. More research is needed not only to devise such potentially better policies but also, to independently test our findings in this paper and validate or invalidate our conclusions regarding the potential benefits of making early bed requests in other hospitals. This is an important avenue of work as it would not only potentially lead to a transformation in the way hospital admissions from the ED are handled but also, instigate a revolution in health-care operations in hospitals with a system-oriented, analytical, and data-driven approach that requires a higher degree of integration between the ED and the main hospital compared with what we have today.

Acknowledgments

The authors would like to thank the associate editor and the two referees, whose comments have helped improve the paper.

Endnotes

¹ Boarding time is defined as the amount of time a patient who is admitted to the hospital spends in the ED bed from the time the admit decision for the patient is given until the patient vacates the ED bed for transfer to the main hospital.

² ED length of stay for a patient is defined as the time between the patient's arrival to the ED until her departure.

References

- Andradottir S, Ayhan H, Down DG (2003) Dynamic server allocation for queueing networks with flexible servers. *Oper. Res.* 51(6):952–968.
- Ang E, Kwasnick S, Bayati M, Plambeck EL, Aratow M (2016) Accurate emergency department wait time prediction. *Manufacturing Service Oper. Management* 18(1):141–156.
- Armony M, Israelit S, Mandelbaum A, Marmor YN, Tseytlin Y, Yom-Tov GB (2015) On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* 5(1):146–194.
- Batt RJ, Terwiesch C (2016) Early task initiation and other load-adaptive mechanisms in the emergency department. *Management Sci.* 63(11):3531–3551.
- Boyle J, Jessup M, Crilly J, Green D, Lind J, Wallis M, Miller P, Fitzgerald G (2012) Predicting emergency department admissions. *Emergency Medicine J.* 29(5):358–365.
- Centers for Disease Control and Prevention (2017) CDC seasonal influenza-associated hospitalizations in the United States. Accessed November 16, 2022, <https://www.cdc.gov/flu/pandemic-resources/tools/flusurge.htm>.
- Chan C, Dong J, Green L (2017a) Queues with time-varying arrivals and inspections with applications to hospital discharge policies. *Oper. Res.* 65(2):469–495.
- Chan CW, Farias VF, Escobar GJ (2017b) The impact of delays on service times in the intensive care unit. *Management Sci.* 63(7):2049–2072.
- Chen W, Linthicum B, Argon NT, Bohrmann T, Lopiano K, Mehrotra A, Travers D, Ziya S (2020) The effects of emergency department crowding on triage and hospital admission decisions. *Amer. J. Emergency Medicine* 38(4):774–779.
- Crilly JL, Boyle J, Jessup M, Wallis M, Lind J, Green D, Fitzgerald G (2015) The implementation and evaluation of the patient admission prediction tool: Assessing its impact on decision-making strategies and patient flow outcomes in 2 Australian hospitals. *Quality Management Health Care* 24(4):169–176.
- Dai JG, Shi P (2017) A two-time-scale approach to time-varying queues in hospital inpatient flow management. *Oper. Res.* 65(2):514–536.
- Dai JG, Shi P (2021) Recent modeling and analytical advances in hospital inpatient flow management. *Production Oper. Management* 30(6):1838–1862.
- Ding Y, Park E, Nagarajan M, Grafsteind E (2019) Patient prioritization in emergency department triage systems: An empirical study of the Canadian Triage and Acuity Scale (CTAS). *Manufacturing Service Oper. Management* 21(4):723–741.
- Diwas SK, Terwiesch C (2017) Benefits of surgical smoothing and spare capacity: An econometric analysis of patient flow. *Production Oper. Management* 26(9):1663–1684.
- Dong J, Shi P, Zheng F, Jin X (2019) Off-service placement in inpatient ward network: Resource pooling versus service slowdown. Preprint, submitted July 18, <http://dx.doi.org/10.2139/ssrn.3306853>.
- Huang J, Carmeli B, Mandelbaum A (2015) Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Oper. Res.* 63(4):892–908.
- Kamali M, Tezcan T, Yildiz O (2019) When to use provider triage at emergency departments. *Management Sci.* 65(3):1003–1019.
- Kuntz L, Mennicken R, Scholtes S (2015) Stress on the ward: Evidence of safety tipping points in hospitals. *Management Sci.* 61(4):754–771.
- Law AM (2007) *Simulation Modeling and Analysis* (McGraw Hill, New York).
- Legros B, Jouini O, Koole G (2018) A uniformization approach for the dynamic control of queueing systems with abandonments. *Oper. Res.* 66(1):200–209.

- Long EF, Mathews KS (2017) The boarding patient: Effects of ICU and hospital occupancy surges on patient flow. *Production Oper. Management* 27(12):2122–2143.
- Mehrotra A, Travers D, Chen W, Lopiano K, Bohrmann T, Argon N, Ziya S, Strickler J, Ring J, Linthicum B (2017) Starting with a clear endpoint: Development of a tool to predict admissions at triage. *Special Issue 2017 Annual Meeting Supplement, Academic Emergency Medicine* 24(S1):S13–S14.
- Peck JS, Benneyan JC, Nightingale DJ, Gaehde SA (2012) Predicting emergency department inpatient admissions to improve same-day patient flow. *Academic Emergency Medicine* 19(9):E1045–E1054.
- Peck JS, Gaehde SA, Nightingale DJ, Gelman DY, Huckins DS, Lemons MF, Dickson EW, Benneyan JC (2013) Generalizability of a simple approach for predicting hospital admission from an emergency department. *Academic Emergency Medicine* 20(11):1156–1163.
- Qiu S, Chinnam RB, Murat A, Batarese B, Neemuchwala H, Jordan W (2015) A cost sensitive inpatient bed reservation approach to reduce emergency department boarding times. *Health Care Management Sci.* 18(1):67–85.
- Saghafian S, Hopp WJ, Van Oyen M, Desmond JS, Kronick SL (2012) Patient streaming as a mechanism for improving responsiveness in emergency departments. *Oper. Res.* 60(5):1080–1097.
- Saghafian S, Hopp WJ, Van Oyen M, Desmond JS, Kronick SL (2014) Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing Service Oper. Management* 16(3):329–345.
- Shi P, Helm JE, Deglise-Hawkinson J, Pan J (2021) Timing it right: Balancing inpatient congestion vs. readmission risk at discharge. *Oper. Res.* 69(6):1842–1865.
- Shi P, Chou MC, Dai JG, Ding D, Sim J (2016) Models and insights for hospital inpatient operations: Time-dependent ed boarding time. *Management Sci.* 62(1):1–28.
- Somanchi S, Adjerid I, Gross R (2022) To predict or not to predict: The case of inpatient admissions from the emergency department. *Production Oper. Management* 31(2):799–818.
- Song H, Tucker AL, Murrell KL (2015) The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Sci.* 61(12):3032–3053.
- Xuang K, Chan C (2016) Using future information to reduce waiting times in the emergency department via diversion. *Manufacturing Service Oper. Management* 18(3):314–331.
- Zayas-Caban G, Xie J, Green LV, Lewis ME (2016) Dynamic control of a tandem system with abandonments. *Queueing Systems* 84(3):279–293.
-
- Wanyi Chen** is a research scientist/decision science consultant at Massachusetts General Hospital. Her expertise is in the areas of health technology assessment and medical decision science using methods including health economics modeling, statistical simulation, cost-effectiveness analysis, and value of information analysis.
- Nilay Tanik Argon** is a professor of statistics and operations research at the University of North Carolina. Her research interests are in stochastic modeling of manufacturing and service systems, queueing systems, healthcare operations, and statistical output analysis for computer simulation.
- Tommy Bohrmann** is a collaborative, entrepreneurial statistician interested in developing and applying analytics to improve healthcare, public health, and the environment. He has extensive experience working with researchers in the areas of medicine, public health, manufacturing, engineering, toxicology, epidemiology, and environmental science.
- Benjamin Linthicum** is a nurse practitioner and adjunct assistant professor of emergency medicine at the University of North Carolina. His interests include improving efficiency of emergency department operations through crossdiscipline collaboration.
- Kenneth Lopiano** is a statistician and the founder and chief executive officer of Roundtable Analytics, Inc.
- Abhishek Mehrotra** is a professor of emergency medicine and the vice chair for strategic initiatives and operations at the University of North Carolina.
- Debbie Travers** is a professor of informatics at Duke University. Her research interests are in emergency department triage, patient throughput, and clinical decision support.
- Serhan Ziya** is a professor of statistics and operations research at the University of North Carolina. His research interests are in service operations, with a focus on healthcare operations, queueing systems, revenue management, and pricing.