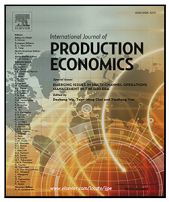
Int. J. Production Economics 231 (2021) 107842

Contents lists available at ScienceDirect

International Journal of Production Economics

journal homepage: www.elsevier.com/locate/ijpe



Proactive coordination of inpatient bed management to reduce emergency department patient boarding

Seung-Yup Lee [a,b,*], Ratna Babu Chinnam [c], Evrim Dalkiran [c], Seth Krupp [d], Michael Nauss [d]

[a] Department of Anesthesiology, School of Medicine, Vanderbilt University Medical Center, 1211 Medical Center Dr, Nashville, TN 37232, USA
[b] Owen Graduate School of Management, Vanderbilt University, 2301 Vanderbilt Place, Nashville, TN 37235, USA
[c] Department of Industrial & Systems Engineering, Wayne State University, 4815 Fourth St, Detroit, MI 48202, USA
[d] Department of Emergency Medicine, Henry Ford Health System, 2799 W. Grand Blvd, Detroit, MI 48202, USA

ARTICLE INFO

Keywords:
Healthcare operations
ED crowding
Proactive coordination
Early task initiation
Predictive analytics
Fork–join queues

ABSTRACT

Emergency departments (EDs) across the world are experiencing severe crowding and prolonged patient wait times for hospital admissions (a.k.a. patient "boarding"). Using data from a major healthcare system, we show that EDs suffer from severe boarding not only due to a high level of hospital inpatient bed occupancy but also due to *reactive* coordination of inpatient bed management activities. To reduce patient boarding, we explore early task initiation for the service network spanning the ED and inpatient units within a hospital. In particular, we investigate the value of predicting ED patient admissions (to be specific, disposition decisions) during the ED caregiving process to proactively initiate downstream tasks for reduced patient boarding. We show that the coordination mechanism can be modeled as a fork–join queueing system. The proposed modeling framework accounts for both imperfect patient disposition predictions and multiple hospital admission sources (in addition to the ED) for inpatient units. We maintain analytical tractability while preserving the complexities of real-world inpatient bed management operations by characterizing the state sets and transition sequences through the Markovian assumption. The proactive inpatient bed allocation scheme can lead to significant reductions in bed allocation delays for ED patients (nearly up to ~ 50%) and does not increase delays for other admission sources. The insights from our model should guide hospital managers in embracing proactive coordination and adaptive workflow technologies enabled by modern health information technology systems and predictive analytics.

1. Introduction

Operations in healthcare facilities entail complex interactions between patients, care providers, and resources. In order to improve the quality and safety of healthcare, hospitals are widely adopting health information technology (HIT) along with lean and six-sigma programs. For instance, by 2015, over 84% of the hospitals in the United States (US) had adopted an electronic health record (EHR) system, a nine-fold increase since 2008 (Henry et al., 2016). While EHR systems increasingly form a critical data backbone for hospital facilities, there is a need for improved *workflow* coordination tools that can enhance situational awareness and facilitate effective management of resources for enhanced efficiency (Kellermann and Jones, 2013). In this paper, we explore *proactive* coordination methods in the form of early task initiation (ETI) for improved operational efficiency of the service network spanning the emergency department (ED) and associated hospital inpatient units (IUs). As a major patient gateway to hospitals and accounting for more than 50% of inpatient admissions in the US (Abelson, 2013), the care quality in the ED significantly affects the overall health outcomes of patients. In particular, growing ED patient "crowding" in recent years has been deemed an international crisis and has received significant public and academic attention (Hoot and Aronsky, 2008; Carter et al., 2014) due to its adverse outcomes, including treatment delays and dissatisfaction (Liu et al., 2003), hospital financial losses (Pines et al., 2011), and harm to staff (Jelinek et al., 2010), to name only a few.

The literature focusing on how to alleviate ED crowding and improve patient flow has been growing. "Output" factors, including prolonged patient admission delays in the ED-to-IU network, a.k.a. "boarding" delays, have been identified as a dominant contributor to severe ED crowding (U.S. GAO, 2003; Fatovich et al., 2005; Van der Vaart

* Corresponding author.
E-mail addresses: seungyup.lee.1@vanderbilt.edu (S.-Y. Lee), ratna.chinnam@wayne.edu (R.B. Chinnam), ey5796@wayne.edu (E. Dalkiran), skrupp1@hfhs.org (S. Krupp), mnauss1@hfhs.org (M. Nauss).

https://doi.org/10.1016/j.ijpe.2020.107842
Received 20 February 2020; Received in revised form 23 May 2020; Accepted 17 June 2020
Available online 22 June 2020

et al., 2011; Shi et al., 2015; Saghafian et al., 2015; Armony et al., 2015). "Boarding" refers to situations where patients who are to be admitted into the hospital are held up in the ED after completing their ED treatment, utilizing expensive resources while they wait for inpatient beds to be prepared and allocated. According to a survey of 1195 US EDs, boarding time accounted for ∼ 37% of the time an admitted patient spent in an ED on average (Augustine, 2016a). Regarding the impact of boarding delays, it has been conservatively estimated that, in addition to compromised health outcomes, EDs experiencing severe boarding delays lose $15,500 of their daily revenue compared to the average ED. This includes both direct losses arising from patients who walk away due to prolonged wait times or are diverted to other hospitals due to the lack of ED beds, as well as losses arising from lost admissions to IUs (Augustine, 2016b). Recognizing the detrimental effects of boarding delays, the Center for Medicare & Medicaid Services (CMS) has required hospitals in the US to report the duration of boarding delays as an ED crowding performance measure since 2014 (CMS and Joint Commission, 2017).

To mitigate boarding, hospital admissions of ED patients can be predicted early during the ED caregiving process and communicated to downstream units so that they can proactively take the necessary steps to reduce transfer delays (Peck et al., 2012, 2014). However, the current practice is to initiate inpatient bed requests and preparation upon confirmation of patient admission at the end of the ED caregiving cycle (Peck et al., 2012), which can potentially lead to extended patient boarding. Thomas et al. (2013) investigated inpatient bed management and advocates proactive bed management through optimized bed–patient assignment decisions for meaningful benefits. Batt and Terwiesch (2017) also discuss the concept of ETI in ED operations. However, unlike their approach, in the coordination strategy we propose, the service providers at the upstream stage (i.e., EDs) and the downstream stages (i.e., IUs and support services) focus on their own tasks while proactively guiding the downstream inpatient bed allocation processes by leveraging future state predictions for ED patients. A prerequisite for proactive inpatient bed management for ED patients is the ability to forecast the trajectories of ED patients. First, predictions of patient admissions and, if they are admitted, the clinically appropriate IUs (a.k.a. "dispositions") must be identified in order to determine the proper IUs for initiating advance bed allocation processes. Second, it is necessary to estimate patients' remaining length of stay (LoS) within the ED in order to initiate bed allocation with a proper lead time. Fortunately, recent advances in predictive analytics methods and tools (including statistical and machine learning methods exploiting patient data from EHR systems) have led to a growing body of literature on effective future state prediction for ED patients (e.g., Qiu et al. 2015, Golmohammadi 2016, Lee et al. 2019, Araz et al. 2019 for admission/disposition predictions and Kocher et al. 2012, Casalino et al. 2014, Chaou et al. 2017 for remaining ED LoS estimation). For the sake of brevity, we note two representative papers concerning the technical advances in the ED patient prediction domain. Lee et al. (2019) built multi-class classification models for progressively predicting disposition decisions for ED patients at different ED caregiving epochs, achieving area under curve values of 0.97, 0.95, 0.89, and 0.84 for the admission classes, i.e., intensive care unit, stepdown/telemetry unit, general unit, and observation unit, respectively. For LoS estimation, Chaou et al. (2017) explored fifteen factors statistically associated with the ED LoS and established the predictive validity of a multivariate accelerated failure time model. However, despite the growing attention being given to the prediction domain, the operationalization of prediction information has not been fully investigated. In addition, the literature reports widely varying prediction performance, depending on the maturity of the employed HIT systems as well as internal practices in making decisions. As a consequence, two significant questions remain unanswered: (1) How can we facilitate proactive bed management and promote its operational efficacy under imperfect predictions? and (2) How can we preserve boarding delays for non-ED admission sources (e.g., referrals, transfers from other hospitals, and transfers from ambulatory surgery center) and the workload of support services while alleviating the delays for ED patients?

We rely on a queueing network representation to model the complex operational interactions between tasks and resources in inpatient bed allocation processes in order to analyze the operational impact of our proposed proactive coordination scheme. Our modeling and analysis allow for quantification of the potential boarding delay reduction as a function of the bed request signal lead-time, the quality of disposition predictions, patient arrival rates from two admission sources, i.e., the ED and non-ED, and bed preparation time. The proposed coordination scheme, which was developed in collaboration with a leading Midwestern US healthcare facility, is practical and can be operationalized in hospitals without requiring significant changes to current clinical and bed management practices. Through this work, we make the following contributions:

1. From a modeling perspective, we develop a scalable and reusable method to analyze parallel (or concurrent) operations with fork–join queue structure models entailing multiple types of tasks and resources and state-dependent transition behaviors. This operational setting combined with the complexity of the operational rules that govern many real-world service systems, renders traditional queueing theory analysis ineffective in answering how a system will behave as a result of interactions among multiple operational parameters. We propose an intuitive yet effective approach for tackling such model settings through the effective representation of generator matrices for state transitions, by hierarchically partitioning the state space under Markovian settings. Without overly simplifying important operational interactions to be analyzed or relying on simulation, the modeling method successfully preserves the complexities of the operations of a system and produces analytical solutions.
2. From an application perspective, the method to analyze fork–join queue structure models allows complex operational rules/requirements for proactive task-resource assignment to be modeled and analyzed. Our approach enables to fully characterize the steady-state distribution of ED-to-IU networks, a relatively complex healthcare service network setting, that operate under real-world bed management rules. By relying on the modeling method, we are able to incorporate multiple important operational parameters into a single model to produce their aggregate impact on the performance measures of the system. To the best of our knowledge, this is the first study to explore the formulation and modeling of fork–join queueing networks that can account for potential errors in work-flow prediction models and relatively complex operational parameters to analyze their joint impact on the system performance.

The key insights resulting from our study include the following: (1) Proactive inpatient bed allocation is a feasible and effective approach particularly during the afternoon and evening hours. This is when EDs are highly congested while unoccupied inpatient beds are available due to a surge in inpatient discharges in the early afternoon. (2) Proactive bed request signals enabled by ED patient disposition predictions can lead to a significant reduction in ED patient boarding when Type-I and Type-II disposition prediction errors are judiciously handled. Finally, (3) because of the distinct trajectories of prediction quality evolvement for ED patients destined to different types of IUs, optimal bed request signal lead times vary depending on the admitting IUs. The insights from our study can guide hospital managers in embracing proactive coordination and adaptive workflow technologies that are enabled by modern HIT systems and predictive analytics.

2. Clinical setting & observations

This section briefly introduces the operations of the ED-to-IU network of a leading academic urban level-1 trauma center by analyzing data gathered over multiple years (May 1, 2014–December 15, 2016), comprising 243,745 ED visits and 41,942 inpatient unit admissions (~17.2% admitted). Upon completing the observation and treatment, an ED physician decides whether the patient should be admitted for further inpatient care or discharged. If admission is necessary, the physician determines the IU that is clinically most appropriate for the patient (i.e., the disposition decision). Once the necessary inpatient bed allocation processes are completed for the patient, the patient is physically transferred to the IU. The two components, i.e., bed allocation delay and patient transfer delay, form boarding delay. Like most EDs, the study hospital's ED suffers from severe boarding delays. During the studied period, a typical admitted ED patient spent 4 h and 39 min receiving care within the ED and an additional 3 h and 2 min of boarding delay after admission approval (median values). The facility is quite representative of many EDs across the US (CMS, 2018) and most likely around the world. With the goal of mitigating boarding delays, we aim to reduce the "bed allocation delay", which is the major portion of boarding delays (with 102.2 min on average at the study hospital) and is prolonged by ineffective and reactive bed allocation operations.

It is worth studying ED patient crowding and related patterns observed within the ED-to-IU network of the study hospital in order to better comprehend the factors influencing boarding delays. Fig. 1 displays ED and IU patient censuses, along with admission and discharge rates by the hour of the day at the study hospital during the data collection period (686 weekdays). Fig. 1(a) shows that the ED suffers from severe crowding (indicated by the dashed line with the secondary y-axis) during the afternoon and evening and is influenced by the increasing levels of patient boarding (indicated by the solid line with the primary y-axis). The shaded area indicates the time interval within each day when the ED suffered from severe crowding, and it can be seen that the high level of boarding did not drop until midnight. Zhou et al. (2012) and George and Evridiki (2015) suggest that there is a significantly increased incidence of serious complications for boarded patients as EDs reach high levels of occupancy. Therefore, the lagged decline of patient boarding levels is detrimental for ED operations, and boarding delays during this time interval are regarded as more harmful to the quality of ED care and should be tightly controlled.

Meanwhile, Fig. 1(b) provides strong evidence that there is a missed opportunity to better coordinate bed allocation processes for many inpatient beds during the ED crowding time interval. The figure plots the probability that there are at least two unoccupied beds (either clean or dirty) in each IU. We conservatively excluded the case of having exactly one unoccupied bed to account for the possibility that a bed has been blocked due to infection concerns about a patient in the same room or is otherwise temporarily unavailable. Contrary to the common belief that IUs are fully occupied during severe ED boarding, there are unoccupied inpatient beds until midnight, when the ED suffers from excessive boarded patients. While the different IUs show their own daily occupancy trends, they share a common pattern of increased availability of unoccupied beds after 2 p.m.

Finally, Fig. 1(c) clearly shows that from noon onward, the number of inpatient beds that become unoccupied increases rapidly and far exceeds the number of beds required for admissions. While this provides evidence for the availability of unoccupied inpatient beds during this time interval, the drastic discharge peak and continuous bed requests in the afternoon and evening bring complications that hamper the effective coordination of inpatient bed allocation processes. Consequently, prolonged admission delays and resulting boarded patients in the ED persist until midnight and even beyond (Fig. 1(a)). A detailed description of the complications is presented in Section 3.

In summary, contrary to common belief, the daily window during which the ED experiences severe congestion with boarded patients occurs when IUs have the largest number of unoccupied beds during the day. Moreover, the unoccupied inpatient beds are not quickly occupied, even when the hospital faces high inpatient bed demand. These patterns are typical of most hospitals across the US, as it is a common practice in the industry that most bed requests are made during the afternoon and evening and most inpatients are discharged around noon after being examined by providers during their morning rounds (Wertheimer et al., 2014). We suggest that even if we take the ED arrival and service rates as given, there are periods during each day when inpatient bed allocations for admitted ED patients can be expedited by means of proactive coordination, without modifying any patient care routines in either the ED or IUs.

3. Proactive coordination of inpatient bed allocation processes

In this section, we take a closer look at the ED-to-IU network operations in the study hospital and propose guiding inpatient bed allocation processes by advance bed request signals. Our approach aims to complete inpatient bed allocation processes near "just-in-time" for ED patient disposition decisions in order to reduce boarding delays.

3.1. Current reactive inpatient bed allocation workflow

Like most hospitals, the study hospital has three main inpatient care units: the general unit (GU), the telemetry unit (TU, also known as the "stepdown" unit), and the intensive care unit (ICU). Each main unit is further categorized into multiple IUs based on specialty (e.g., Internal Medicine Unit in the GU and Cardiovascular Surgery ICU in the ICU). It is worth noting that there are groups of IUs within a main IU that share the common features of inpatient beds and accessories. This in

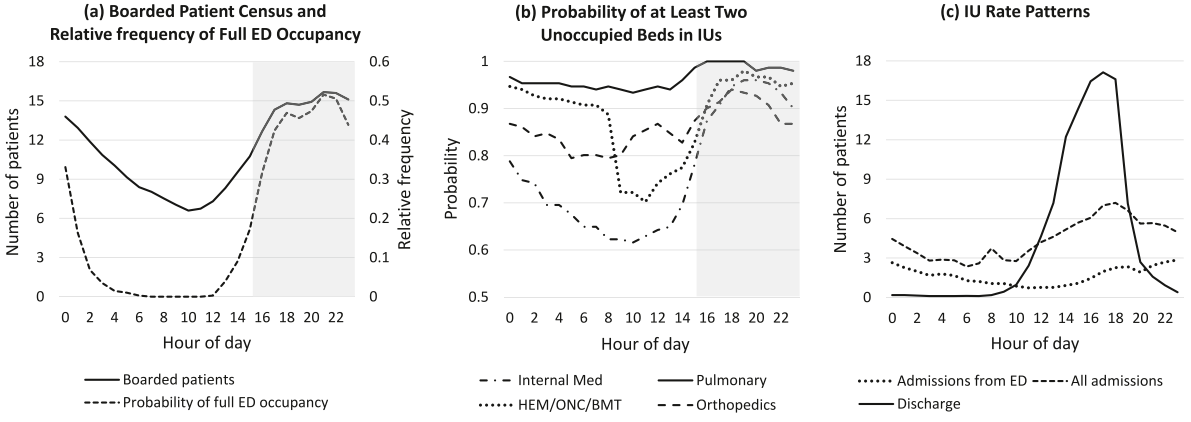


Fig. 1. ED and IU patient censuses and flow rates by hour of day.

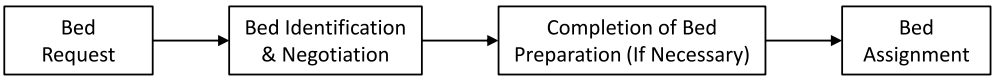


Fig. 2. Typical inpatient bed allocation process in the ED-to-IU network.

turn constrains patient "overflow" to occur within an IU group , and beds can be shared only within the group. The study hospital has environmental services (EVS) teams that are in charge of cleaning inpatient beds as well as other essential tasks including area decontamination, hygiene management, and managing linen/paper and other supplies for the area. Generally, in a large hospital, each EVS worker is assigned to an IU group to reduce unnecessary travel between assignments and to ensure that patient safety and hygiene are tightly managed within the assigned IU group. Such a dedicated system grants more ownership and responsibility over the tasks to the workers, allowing them to strategically adjust their work routines to handle workload more efficiently (Song et al., 2015).

The general bed allocation process within the ED-to-IU network is illustrated in Fig. 2. While the bed allocation process shown in this figure is common to all patients who are being hospitalized, bed preparation becomes dependent on the admission flow when there are operational complications that force the bed preparation process to be conditional on bed requests. To be specific, uncertainty regarding bed demand presents challenges for making the best use of the limited EVS staff. For example, upon completing a current EVS task, should the staff member clean an adjacent dirty bed in the same area (for which no patient is currently waiting), give priority to cleaning a dirty bed in a distant unit (incurring non-value-added travel time), or attend to other duties that promote safety (such as area decontamination and hygiene management)? This complication in the operations becomes more prominent when EVS teams become busy to respond to service needs and unclear in prioritizing their tasks. As a result, EVS teams prepare inpatient beds in accordance with the realized bed demand, investing their workforce in certain and immediate EVS needs. While this strategy helps an EVS team prevent the misallocation of its staff members for taking care of urgent tasks, it can generate prolonged bed allocation delays for admitted ED patients. Therefore, bed allocation delays can include some or all of the components of bed preparation delay, which consist of the time bed managers need for communicating with EVS servers in order to deploy them, the time needed to finish duties associated with any prior assignments, the travel time taken by servers to move to the assigned bed location, and bed cleaning delays.

Given that inpatient bed preparation is dependent on the admission flow during the time window of interest, our proposed approach proactively prioritizes the task of inpatient bed preparation for those EVS servers in IUs that face immediate or projected demand from the ED. We demonstrate that if EVS servers were to receive advance bed request signals for likely admissions to each IU with adequate and reasonable lead times, they would be able to manage their work in a proactive manner to reduce ED boarding and, in turn, ED crowding. The hospital already employs communication tools and platforms (e.g., the EPIC EHR system combined with pagers for EVS servers and bed managers) that can be leveraged for operationalizing the proactive bed request signals.

3.2. Scheme for proactive inpatient bed allocation

Fig. 3 illustrates the ETI strategy for enabling proactive bed allocation by exploiting EHR data. The key assumption is that once a patient enters the ED and begins to undergo triage, testing, and treatment, there is adequate and increasing information about the patient within the EHR, including the patient's health history, to allow more reliable predictions of ED disposition decisions ahead of the actual disposition decisions (Golmohammadi, 2016; Araz et al., 2019). This will enable the ED to proactively signal the relevant IU regarding an impending admission and need for a bed.

For implementing our proposed proactive inpatient bed allocation strategy, Fig. 4 depicts the underlying concurrent processes of two different server systems for an IU group, say IU ω (henceforth, the ED-to-IU$_\omega$ network). The fundamental idea of the proposed scheme is that immediately after sending a bed request signal to IU ω, the work flow "forks" into two different routes: (1) the remaining ED treatment processes for the patient (server s_ω^1) and (2) the bed allocation/preparation process (server s_ω^2). A bed assignment is instantly made at the "join" server s_ω^3 when both a patient and a bed are present in their respective queues for pairing at server s_ω^3. For managerial insights, our fork–join queue modeling approach incorporates the following main considerations:

1. *Impact of Bed Request Signal Lead Time*: In our proposed proactive coordination scheme, IU bed managers and EVS staff initiate bed allocation processes in a first-in-first-out (FIFO) manner based on the proactive bed requests they receive from the ED. This in turn enables server s_ω^2 to manage bed preparation in exactly the same way that it handles bed preparation tasks in the reactive

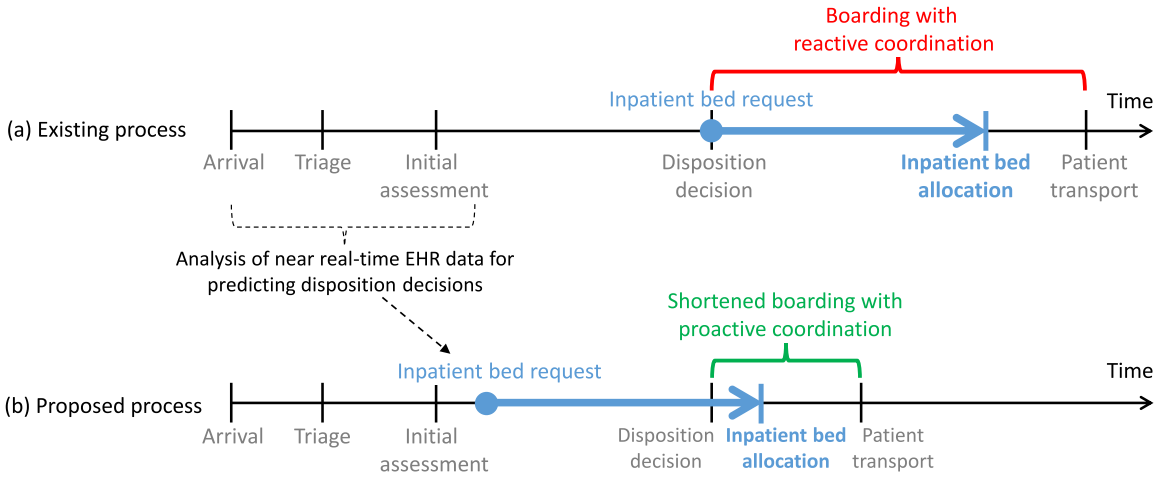


Fig. 3. Proactive inpatient bed allocation strategy for reducing boarding delay.

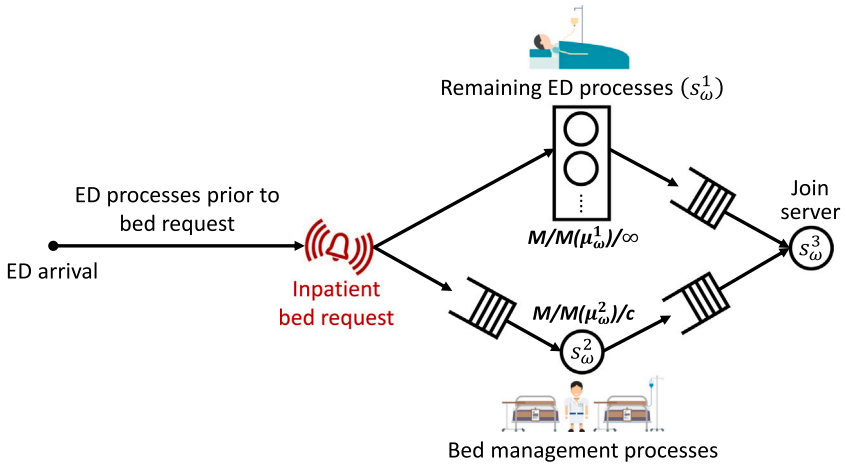


Fig. 4. Preliminary scheme for proactive inpatient bed request and allocation.

case, with the only difference being that the requests from the ED are now proactive. Under this setting, our modeling can provide insights into the impact of the extent of proactivity in bed allocation (i.e., the bed request signal lead time) on reducing boarding delays. While it may be desirable to send bed requests for patients at the earliest stages of the ED treatment to minimize bed allocation delays, in reality, accurate predictions can only be made once the patient undergoes adequate monitoring, testing, and treatment. This necessarily limits the bed request signal lead time to attain further reductions in bed allocation delays by promoting higher prediction quality. Whereas quality predictions can be achieved by decreasing the signal lead time, the original operational benefit from practicing proactive coordination is thereby compromised. This trade-off between the extent of proactivity and the quality of predictions suggests that there may be an optimal signal lead time that minimizes bed allocation delays. Moreover, since the prediction quality for ED patients destined for different types of IUs exhibits different growth trajectories throughout the ED caregiving processes, optimal signal lead times will vary depending on the admitting IUs.

2. *Uncertainty about the Remaining ED Service Time*: The bed request signal entails specifying the timing of ED treatment completion, which is intrinsically uncertain. In our modeling approach, we represent this uncertainty regarding the remaining ED treatment by an exponentially distributed service time for server s^1_ω. Since the coefficient of variation for an exponentially distributed random variable is unity, we implicitly and reasonably model the remaining service time in s^1_ω with a variance that increases along with the mean. That is, the earlier we send a bed request signal for a patient upon arrival at the ED, the higher the variance of service time at server s^1_ω. This is consistent with the observation that as patients go through more ED treatment, the variation in their remaining ED service time decreases.
3. *Allowing Overflow Only within an IU Group*: While hospitals group beds for distinct IUs based on their particular features, recommending against "off-unit placement" of patients, the beds are regarded as indistinguishable across subunits within an IU group, which allows patient overflow within an IU group (if there are no beds available in the preferred specialty subunit). Given the practice, there is no need to match individual ED patients to specific beds within IU ω; a FIFO treatment is appropriate. Hence, the bed assignment process in server s^3_ω is modeled as a FIFO process.
4. *Significant Number of Unoccupied Inpatient Beds:* As demonstrated in Fig. 1, a significant number of unoccupied inpatient beds are generated during the peak discharge period. This results in unoccupied beds waiting for preparation and assignment on most days during the afternoon and evening periods. Moreover, the routine patient overflow occurring within an IU group further increases the availability of unoccupied inpatient beds during this period. For instance, the mean numbers of unoccupied beds within an area of GU (on the same floor) that frequently share resources were 12.05, 15.98, and 14.09 at 2 p.m., 7 p.m., and midnight, respectively, indicating the sufficient availability of unoccupied inpatient beds until midnight. Hence, we have reason to assume that there are unoccupied inpatient beds upon bed requests during the ED crowding period (the main focus interval for proactive coordination).

Presuming that the set of ED processes that occur prior to the bed request signals operates as a stable system (i.e., service rate > arrival rate), it is reasonable to assume that the bed request signals follow the same Poisson process arriving into the ED. While the ED arrival process is known to be a time-varying Poisson process, in this paper we focus on modeling the complex network behaviors to evaluate the impact of early bed request signals and patient classification errors in a homogeneous Poisson arrival setting. The remaining ED processes are represented as an ∞-server system (server s^1_ω). The justification for the infinite number of servers is that the ED patients in server s^1_ω are already under the care of ED providers, so there is no queue upon arrival at the server. Therefore, service time in server s^1_ω represents the bed request signal lead time that primarily implies the level of proactivity. In addition, the interdeparture time of server s^2_ω in the reactive case approximates exponential distributions in the hospital's representative IUs (see Online Appendix OA.1). Based on this observation, we assume that server s^2_ω's service time follows an exponential distribution.

The proposed fork–join queue network is unique in that the configurations of the servers in the two-server system are dissimilar, and so are their service times. Due to service time uncertainty and the differences in the two-server system, two queues will be formed for server s^3_ω, which results in wait times. This baseline model is modified significantly in Section 4 by incorporating the important aspects of this study, i.e., multiple arrival sources, prediction errors, and queue management rules, which make the network setting and operations mimic a real-world setting. The main objective of this study is to model and analyze the expected wait times in each of the join server s^3_ω's queues in accordance with several modeling parameters in order to evaluate the operational impacts of proactive coordination within the ED-to-IU$_\omega$ network.

4. Modeling framework

This section introduces additional modeling details that further shape the baseline model to account for the realities of ED-to-IU$_\omega$ network operations. Unlike the traditional fork–join queueing structures,

we assume that there are multiple types of tasks and resources (i.e., heterogeneity), and only specific types of tasks are the beneficiaries of proactive resource allocation for operational purposes (i.e., selectivity). Moreover, because we rely on imperfect predictions for individual patients about expected task types and flows (prior to learning their true task types/flows), the modeling framework should be able to adequately address false predictions (revealed with some delay) by fulfilling a set of task-resource assignment rules designed to control prediction error cases (to be discussed in detail in Section 4.2). Even though the modeling considerations result in complex state-dependent transition behaviors, we aim to explicitly model the details to fully characterize the steady-state distribution by investigating the operational characteristics of the states.

4.1. Types of patients and beds in the proposed network

IU ω serves patients being admitted from both the ED and other non-ED admission sources (including outpatients, transfers from other hospitals, and transfers from different departments within the hospital) that come with care needs that fit medical specialty in the IU group. Without loss of generality, we treat patients being admitted to IU ω from all sources other than the ED as a single type of patient, i.e., $\widetilde{E}_\omega$ patients. This forms two patient demand types for IU ω: E_ω (patients admitted from the ED to IU ω) and $\widetilde{E}_\omega$ patients, while creating the single queue at IU ω. If not handled effectively, a bed readied for an E_ω patient might be occupied by an $\widetilde{E}_\omega$ patient, compromising the operational benefit for E_ω patients. To avoid such an unintended situation, we propose proactive bed allocation that *reserves* IU beds for the E_ω class of patients (that is, not for individual patients but rather for the class as a whole). Through these reservations, beds prepared in response to advance ED requests will be dedicated to E_ω patients and can only be occupied by them. In addition, E_ω patients cannot take beds being prepared for $\widetilde{E}_\omega$ patients, which ensures that our approach will not compromise the efficiency of bed allocation for $\widetilde{E}_\omega$ patients. Given the primary aim of this study, we limit proactive bed allocations to ED patients alone (i.e., selectivity).

Patient disposition predictions are error prone due to the clinical uncertainty about ED patients, and these predictions can therefore be a source of bed request errors owing to disagreements between disposition decision predictions and true disposition decisions. Specifically, from the viewpoint of IU ω, the predicted disposition for a patient will be either positive, meaning "will be admitted to ω", or negative, meaning "will not be admitted to ω". Depending on the predicted and true dispositions for a patient with respect to IU ω, a prediction will be either correct or incorrect, and these patients must be treated in different ways in patient–bed assignment within the ED-to-IU$_\omega$ network.

Fig. 5 provides a queueing network representation of the ED-to-IU$_\omega$ network, factoring in the classification errors as well as other admission sources. An advantage of the representation in Fig. 5 is that it can explicitly specify all of the classification instances. A summary of the model notation is presented in Table 1. Let a type E_ω^{AB} patient be an ED patient with predicted disposition $A \in \{P, N\}$ and true disposition $B \in \{\emptyset, P, N\}$ for IU ω, where P means "positive", N means "negative", and $\emptyset$ indicates that the disposition decision is yet to be made. We utilize boldface $\boldsymbol{X}$ to represent the set of objects (patients/beds) of type X.

The arrival of ED patients in the network is modeled as a Poisson process with rate λ^E. An ED patient turns into either an $E_\omega^{P\emptyset}$ (*positively* predicted for IU ω) patient with probability p_ω or an $E_\omega^{N\emptyset}$ (*negatively* predicted for IU ω) patient with probability $1 - p_\omega$. An $E_\omega^{P\emptyset}$ patient then becomes either a E_ω^{PP} patient with probability q_ω or an E_ω^{PN} (Type-I error) patient with probability $1 - q_\omega$, depending on the true disposition for the patient with respect to IU ω. Also, depending on the true disposition, an $E_\omega^{N\emptyset}$ patient becomes either an E_ω^{NN} patient (with probability $1-r_\omega$) or an E_ω^{NP} patient (Type-II error with probability r_ω). Subsequently, only E_ω^{PP} and E_ω^{NP} patients enter server s_ω^3 to occupy a bed ($\boldsymbol{E}_\omega = \boldsymbol{E}_\omega^{PP} \cup \boldsymbol{E}_\omega^{NP}$) in IU ω. Since inpatient beds are assumed to be indistinguishable/homogeneous within IU ω, a bed that has been reserved upon receiving a bed request made for an $E_\omega^{P\emptyset}$ patient can be assigned to *any* E_ω^{PP} or E_ω^{NP} patient in a FIFO manner. Table 2 summarizes the seven types of patients that can exist and interact within the network, corresponding to the predicted and true dispositions for patients with respect to IU ω.

The errors in classification predictions generate not only different types of patients but also different types of beds. For instance, if an $E_\omega^{P\emptyset}$ patient turns out to be an E_ω^{PN} patient for whom a bed has already been prepared and there are no ED patients within the ED-to-IU$_\omega$ network, the bed should be given to a $\widetilde{E}_\omega$ patient waiting for a bed. This situation therefore requires defining an activity that cancels the bed reservation and releases the bed so that it can be occupied by anyone. We call this action a "release". Since prediction errors can alter the way that beds are assigned to patients, it is necessary to define different types of beds to describe the interactions among patients and beds in different situations. We define six types of beds as shown in Table 3, showing how patient–bed assignments are made according to their types in servers s_ω^2 and s_ω^3, respectively.

In this study, we focus on obtaining the steady-state probability distribution of the numbers of each type of patient and bed in the network, modeling the interactions among the entities depending on predicted and true dispositions. It is worth noting that the properties of Burke's output theorem enable reducing the complexity of the problem (Burke, 1956). First, the arrival processes for $E_\omega^{P\emptyset}$ and $E_\omega^{N\emptyset}$ patients are independent, owing to the property of Poisson processes. Therefore, the departure processes for $E_\omega^{N\emptyset}$ patients in server s_ω^1 can also be analyzed independently. Moreover, the number of $E_\omega^{N\emptyset}$ patients in server s_ω^1 is not of interest since we pay attention to analyzing the queue lengths for servers s_ω^2 and s_ω^3. It is only when an $E_\omega^{N\emptyset}$ patient turns into an E_ω^{NP} entering server s_ω^3 that the queue lengths for servers s_ω^2 and s_ω^3 can be affected. According to Burke's output theorem, the departure of $E_\omega^{N\emptyset}$ patients from server system s_ω^1 in the steady state is also a Poisson process with the rate $(1 - p_\omega)\lambda^E$. Among the departures at rate $(1 - p_\omega)\lambda^E$, $(1 - p_\omega)r_\omega\lambda^E$ patients proceed to server s_ω^3 and then become E_ω^{NP} patients. Therefore, without explicitly tracking the behavior of $E_\omega^{N\emptyset}$ patients in server s_ω^1, we can achieve a complete model to obtain analytical solutions for the steady-state probability distribution of queue lengths at server s_ω^3.

Based on the designed fork–join queue structure, we introduce six activities that trigger a change within the network in Fig. 5: (1) arrival of an $E_\omega^{P\emptyset}$ patient at rate $p_\omega\lambda^E$, (2) arrival of an E_ω^{NP} patient at rate $(1 - p_\omega)r_\omega\lambda^E$, (3) arrival of an $\widetilde{E}_\omega$ patient at rate $\lambda_\omega^{\widetilde{E}}$, (4) completion of the remaining ED processes for an E_ω^{PP} patient at rate $iq_\omega\mu_\omega^1$, where i is the number of $E_\omega^{P\emptyset}$ patients processed by server s_ω^1, (5) completion of the remaining ED processes for an E_ω^{PN} patient at rate $i(1 - q_\omega)\mu_\omega^1$, where i is the number of $E_\omega^{P\emptyset}$ patients processed by server s_ω^1, and (6) completion of the preparation of a bed at server s_ω^2 at rate μ_ω^2. Rather than assuming that all servers and queues respond identically to each activity (which is typical in fork–join queue structures), our model entails more realistic – that is, state-dependent – transition behaviors corresponding to each of the activities for handling prediction errors and the multiple types of patients and beds. In the following section, we present and discuss bed reservation rules that have been established for governing the proactive bed management by mimicking bed assignment mechanisms in practice.

4.2. Inpatient bed reservation strategy and network behaviors

In this section, we propose proactive bed preparation/allocation rules that handle the network cases resulting from Type-I and Type-II disposition prediction errors. To guarantee a maximized positive influence of advance bed request signals in reducing ED patient boarding without compromising the bed allocation efficiency for non-ED patients or the bed preparation efforts of server s_ω^2, we design rules

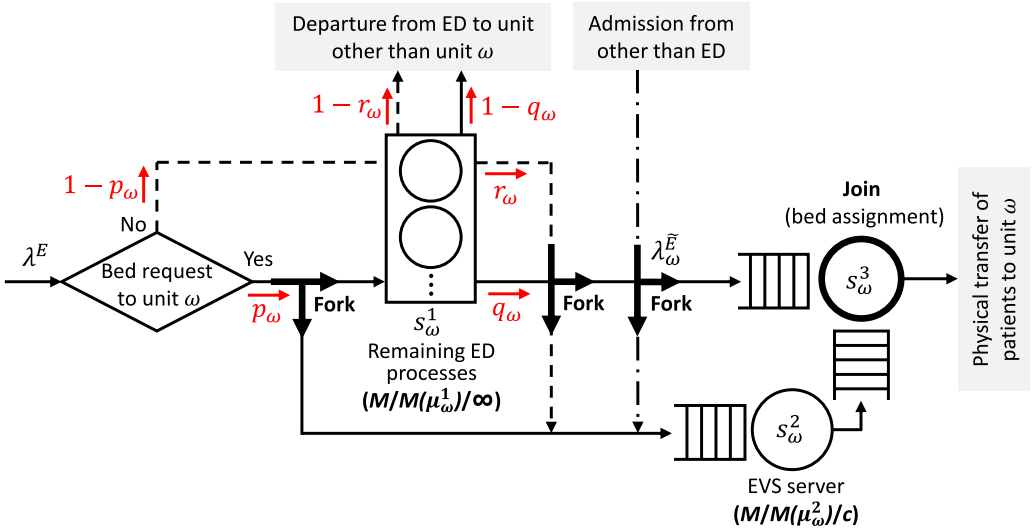


Fig. 5. Fork–Join queue structure for proactive inpatient bed allocation.

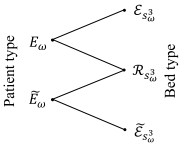
Table 1
Model notation.

Notation	Explanation
p_ω	Probability of sending a bed request to unit ω (exponentially distributed) $1/\mu^1_\omega$ time prior to disposition = P(prediction in regard to IU ω = positive).
q_ω	Probability that an $E^{P\emptyset}_\omega$ patient becomes an E^{PP}_ω patient = P(disposition in regard to IU ω = positive | prediction in regard to IU ω = positive).
r_ω	Probability that an $E^{N\emptyset}_\omega$ patient becomes an E^{NP}_ω patient = P(disposition in regard to IU ω = positive | prediction in regard to IU ω = negative).
u_ω	Proportion of admissions from the ED among all admissions at unit ω
λ^E	Arrival rate of ED patients to the bed request decision node
$\lambda^{\widetilde{E}}_\omega$	Arrival rate of $\widetilde{E}_\omega$ patients at unit ω
$1/\mu^1_\omega$	Bed request signal lead-time for unit ω (service time at server s^1_ω)
μ^2_ω	Bed preparation rate at unit ω (service rate at server s^2_ω)
Z_ω	Admission probability from the ED to unit $\omega = p_\omega q_\omega + (1-p_\omega) r_\omega$.
$\pi_{(\theta)}$	Steady-state probability for state θ
$\mathrm{E}(L^E_\omega)$	Expected number of ED patients awaiting inpatient beds at unit ω
$\mathrm{E}(L^{\widetilde{E}}_\omega)$	Expected number of non-ED patients awaiting inpatient beds at unit ω
$\mathrm{E}(L^P_\omega)$	Expected number of patients awaiting inpatient beds at unit ω $(= \mathrm{E}(L^E_\omega) + \mathrm{E}(L^{\widetilde{E}}_\omega))$
$\mathrm{E}(L^B_\omega)$	Expected number of beds at server s^3_ω
$\mathrm{E}(W^E_\omega)$	Expected bed allocation delay for ED patients admitted at unit ω
$\mathrm{E}(W^{\widetilde{E}}_\omega)$	Expected bed allocation delay for non-ED patients admitted at unit ω
Extensions (key classification performance measures)	
σ^{PPV}_ω	Positive predictive value (PPV, precision) of disposition prediction at unit $\omega = q_\omega$.
σ^{TPR}_ω	True positive rate (TPR, sensitivity) of disposition prediction at unit $\omega = p_\omega q_\omega / Z_\omega$.
σ^{FPR}_ω	False positive rate (FPR) of disposition prediction at unit $\omega = p_\omega(1-q_\omega)/(1-Z_\omega)$.

Table 2
Patient types within ED-to-IU$_\omega$ network.

Admission source	Predicted disposition	True disposition	Patient type	Classification performance	Patient flow volume
ED	Positive	Not made yet	$E^{P\emptyset}_\omega$	Not known yet	$p_\omega \lambda^E_\omega$
		Positive	E^{PP}_ω	True positive	$p_\omega q_\omega \lambda^E_\omega$
		Negative	E^{PN}_ω	False positive (Type-I error)	$p_\omega(1-q_\omega)\lambda^E_\omega$
	Negative	Not made yet	$E^{N\emptyset}_\omega$	Not known yet	$(1-p_\omega)\lambda^E_\omega$
		Positive	E^{NP}_ω	False negative (Type-II error)	$(1-p_\omega) r_\omega \lambda^E_\omega$
		Negative	E^{NN}_ω	True negative	$(1-p_\omega)(1-r_\omega)\lambda^E_\omega$
Non-ED	Not applicable	Positive	$\widetilde{E}_\omega$	Not applicable	$\lambda^{\widetilde{E}}_\omega$

Table 3
Bed types within ED-to-IU$_\omega$ network.

Initially prepared for	Released?	Current location	Bed type notation	Patient–bed match at join server s^3_ω
ED patients	No	s^2_ω	$\mathcal{E}^2_\omega$	
		s^3_ω	$\mathcal{E}^3_\omega$	
	Yes	s^2_ω	$\mathcal{R}^2_\omega$	
		s^3_ω	$\mathcal{R}^3_\omega$	
Other patients (non-ED patients)	Not applicable	s^2_ω	$\widetilde{\mathcal{E}}^2_\omega$	
		s^3_ω	$\widetilde{\mathcal{E}}^3_\omega$	

that strategically govern bed reservations, releases, and cancellations according to the beds' current demand and preparation status. The rules are as follows:

(R1) *Selectivity:* Proactive bed requests are made only for $E_\omega^{P\emptyset}$ patients.

(R2) *Heterogeneity:* ED and non-ED patients are served on a FIFO basis at the join server s_ω^3, but according to the prepared bed types. An ED patient can only take a bed that is either reserved for ED patients or released, and a non-ED patient can only take a bed that is not reserved for ED patients (as graphically represented in Table 3).

(R3) *Error tolerance*: The operational rules that handle prediction error cases are summarized in Fig. 6, where we use the cardinality notation $|\cdot|$ to represent the number of each type of patient and bed and c to denote the number of servers in s_ω^2. Once a false positive is revealed, there is one more bed reserved than are necessary in the system. In this case, one of the reserved beds is released or removed according to the bed preparation/allocation state at the given time (Fig. 6(a)). On the other hand, once a false negative arrives, fewer beds than are necessary have been prepared for ED patients. In this case, either an $\mathcal{E}_\omega^2$ bed is placed in the queue or a released bed is again reserved for ED patients (Fig. 6(b)).

(R4) *Correspondence*: If the handling of prediction errors based on R1, R2, and R3 results in any redundant beds in the queue for server s_ω^2, according to the patient–bed match (Table 3) within the network, the redundant bed in the queue for server s_ω^2 is removed.

(R5) *Successive update*: A chain of patient–bed assignments is successively made after an activity until no additional change is required in the network by rules R1–R4.

We provide an example in which the proposed bed management rules are applied to control bed allocation when a false positive occurs. Consider Fig. 7, where the black and white circles represent ED patients and non-ED patients, respectively, and the black and white squares represent beds that have been prepared or are being prepared for ED patients and non-ED patients, respectively. The gray square represents a released bed. Initially there are one $E_\omega^{P\emptyset}$ patient, two $\widetilde{E}_\omega$ patients, two $\widetilde{\mathcal{E}}_\omega^2$ beds, and one $\mathcal{E}_\omega^3$ bed (Fig. 7(a)). As soon as a false positive is revealed after the true disposition for the $E_\omega^{P\emptyset}$ patient in server s_ω^1, the following events occur in the specified order. As the E_ω^{PN} patient leaves the network, the $\mathcal{E}_\omega^3$ bed is released and becomes an $\mathcal{R}_\omega^3$ bed in accordance with R3 (Fig. 7(b)). Then one $\widetilde{E}_\omega$ patient and the $\mathcal{R}_\omega^3$ bed are joined at server s_ω^3 according to R2, and the $\widetilde{\mathcal{E}}_\omega^2$ bed in the queue for server s_ω^2 is removed in accordance with R4. Eventually, one $\widetilde{E}_\omega$ patient and one $\widetilde{\mathcal{E}}_\omega^2$ bed remain in the network (Fig. 7(c)).

The rules were designed based on prevalent bed allocation practices in the industry, but have been modified to effectively deal with prediction errors in such a way that the list of beds to be prepared is updated based on true dispositions. The rules proactively match the bed supply to the bed demand based on advance bed requests while removing a redundant bed from the queue or releasing it, as indicated in Fig. 6. This approach reduces unnecessary bed wait times while controlling the possible wastage of the workforce incurred by erroneous predictions and does not change the current EVS practice or bed demands other than providing proactive signals to EVS servers. However, the inclusion of prediction information in the network and the treatment of prediction errors elevate the complexity of the queueing network to be analyzed. In the following section, we discuss our approach to representing the network and obtain analytical solutions to evaluate the operational impact of our proposed proactive bed allocation strategy.

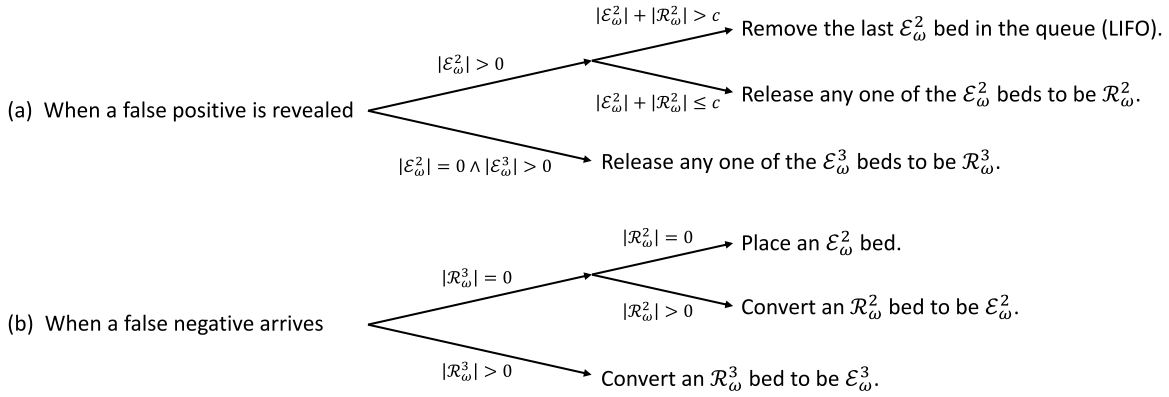


Fig. 6. Inpatient bed reservation rules in case of prediction errors.

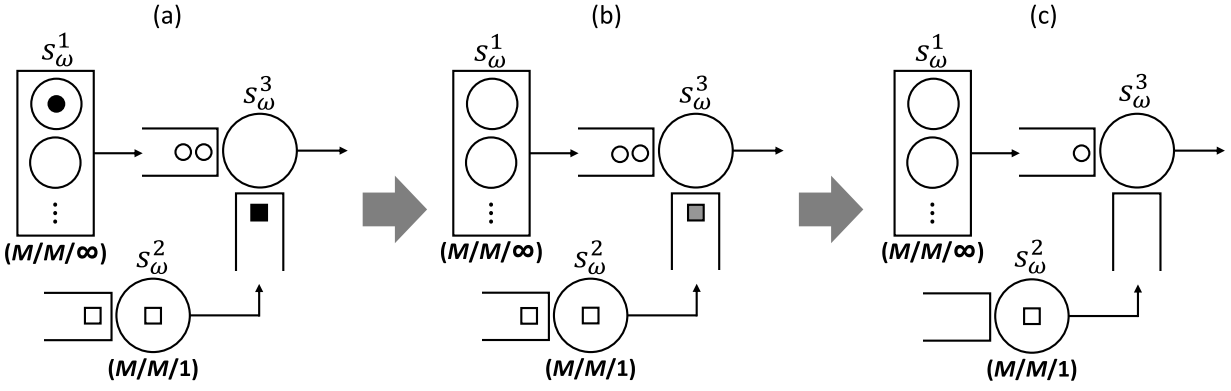


Fig. 7. ED-to-IU$_\omega$ network behaviors responding to a false positive case according to bed reservation rules.

4.3. Representation of the ED-to-IU network state space

In this section we translate the rules and relationships among the different types of patients and beds in the ED-to-IU$_\omega$ network into state-dependent transitions among states representing the network, with the aim of obtaining the steady-state probability distribution of the system. Considering the complexity of the problem generated by the multiple types of entities and state-dependent transitions, one alternative would be to use simulation, but this could result in limited insights. Instead, we propose representing the network as a trivariate Markov process by imposing constraints that categorize the entire state space into manageable subsets.

We represent the proposed ED-to-IU$_\omega$ network as a continuous-time Markov process on the state space $\{(i,\ \boldsymbol{j},\ k) : 0 \le i \le N,\ 0 \le |\boldsymbol{j}| \le N,\ -N \le k \le N\}$, where i is the number of $E_\omega^{P\emptyset}$ patients, $\boldsymbol{j}$ is a string that stores the order of the different types of beds being prepared in server s_ω^2, and k is the number of either boarded ED patients or beds queued at server s_ω^3, depending on the sign of k. In the trivariate representation, we omit the subscript ω for conciseness. A string $\boldsymbol{j}$ is a sequence of two elements, $\widehat{\mathcal{E}}$ and $\widetilde{\mathcal{E}}$, that represent a bed available for ED patients (i.e., either an $\mathcal{E}_\omega^2$ or an $\mathcal{R}_\omega^2$ bed) and a bed unavailable for ED patients (i.e., an $\widetilde{\mathcal{E}}_\omega^2$ bed), respectively. Based on the definition of the string $\boldsymbol{j}$, we omit the superscript for beds for clarity of expression. Let $|\boldsymbol{j}|$ denote the number of elements in the string $\boldsymbol{j}$, and let $|\boldsymbol{j}_{\widehat{\mathcal{E}}}|$ and $|\boldsymbol{j}_{\widetilde{\mathcal{E}}}|$ represent the numbers of $\widehat{\mathcal{E}}$ and $\widetilde{\mathcal{E}}$ elements, respectively, in the string. If $k > 0$, k represents the number of boarded ED patients waiting for beds, and if $k < 0$, $|k|$ is the sum of the numbers of $\mathcal{E}_\omega^3$ and $\mathcal{R}_\omega^3$ beds ready to be occupied (i.e., $k^+ = |\boldsymbol{E}_\omega^{PP}| + |\boldsymbol{E}_\omega^{NP}|$ and $k^- = |\boldsymbol{\mathcal{E}}_\omega^3| + |\boldsymbol{\mathcal{R}}_\omega^3|$).

$\mathcal{E}_\omega^2, \mathcal{E}_\omega^3, \mathcal{R}_\omega^2, \mathcal{R}_\omega^3$ are not explicitly specified by the trivariate state space representation. However, their values can be easily derived, as shown in Table 4. To this end, we partition the state space based on the value of $i + k$, i.e., $i + k = |\boldsymbol{j}_{\widehat{\mathcal{E}}}|$ (no released bed in the system) or $i + k < |\boldsymbol{j}_{\widehat{\mathcal{E}}}|$ (released beds are present). For the first case, we have $|\boldsymbol{\mathcal{R}}_\omega^2| = |\boldsymbol{\mathcal{R}}_\omega^3| = 0$, and the number of ED patients who need beds and the number of beds that are being prepared or are already prepared for ED patients are balanced, i.e., $|\boldsymbol{E}_\omega^{P\emptyset}| + |\boldsymbol{E}_\omega^{PP}| + |\boldsymbol{E}_\omega^{NP}| = |\boldsymbol{\mathcal{E}}_\omega^2| + |\boldsymbol{\mathcal{E}}_\omega^3|$. Hence, $|\boldsymbol{\mathcal{E}}_\omega^2| = |\boldsymbol{j}_{\widehat{\mathcal{E}}}|$ and $|\boldsymbol{\mathcal{E}}_\omega^3| = k^-$. For the second case, released beds are present in the system, and $|\boldsymbol{j}_{\widehat{\mathcal{E}}}| - (i + k)$ is the number of these beds. Note that, by rule R4, preparation of a redundant bed is canceled when the preparation has not yet started. Therefore, Eq. (1) holds true for $M/M/\infty - M/M/c$:

$$i + k < |\boldsymbol{j}_{\widehat{\mathcal{E}}}| \ \rightarrow\ |\boldsymbol{j}_{\widehat{\mathcal{E}}}| < c. \tag{1}$$

The locations of the released beds are determined based on the value of $i + k$. If $i + k \ge 0$, all of the released beds are in server s_ω^2, i.e., $|\boldsymbol{\mathcal{R}}_\omega^2| = |\boldsymbol{j}_{\widehat{\mathcal{E}}}| - (i + k)$ and $|\boldsymbol{\mathcal{R}}_\omega^3| = 0$. Accordingly, $|\boldsymbol{\mathcal{E}}_\omega^2| = i + k$ and $|\boldsymbol{\mathcal{E}}_\omega^3| = k^-$. If $i + k < 0$, then we have $|\boldsymbol{j}_{\widehat{\mathcal{E}}}|$ released beds in server s_ω^2 and $-(i + k)$ released beds in server s_ω^3. Thus, $|\boldsymbol{\mathcal{E}}_\omega^2| = 0$ and $|\boldsymbol{\mathcal{E}}_\omega^3| = i$. By the definitions of non-ED beds and patients, we always have $|\boldsymbol{j}_{\widetilde{\mathcal{E}}}| = |\widetilde{\boldsymbol{\mathcal{E}}}_\omega^2| = |\widetilde{\boldsymbol{E}}_\omega|$, since a bed request for an $\widetilde{E}_\omega$ patient is made only when the patient arrives. This also implies that $|\widetilde{\boldsymbol{\mathcal{E}}}_\omega^3| = 0$.

We have discussed how the seven types of patients and six types of beds are interrelated and can be represented within the trivariate Markov process structure. However, due to the complexity of the state space structure and the state-dependent transition behaviors, simple lexicographic ordering of states does not work. For this reason, we introduce methods that can categorize the entire state space into subsets of states so that transitions can be defined between and within different subsets of states. We define (1) "primary sets" of states $H(n)$, based on the number of bed requests remaining in the network, (2) "secondary sets", based on the sign of k within a primary set, and (3) "tertiary sets", based on the availability of different types of beds within a secondary set. We then define transition patterns across the subsets in order to find an analytical solution to the steady-state distribution by solving global balance equations.

Table 4
Trivariate Expression for Numbers of Entities.

Entity	No released beds		Released beds present		
	$k \ge 0$	$k < 0$	$k \ge 0$	$-i \le k < 0$	$k < -i$
$\lvert\boldsymbol{E}_\omega^{PP}\rvert + \lvert\boldsymbol{E}_\omega^{NP}\rvert$	k	0	k	0	0
$\lvert\boldsymbol{\mathcal{E}}_\omega^2\rvert$	$i + k$	$i + k$	$i + k$	$i + k$	0
$\lvert\boldsymbol{\mathcal{E}}_\omega^3\rvert$	0	$-k$	0	$-k$	i
$\lvert\boldsymbol{\mathcal{R}}_\omega^2\rvert$	0	0	$\lvert\boldsymbol{j}_{\widehat{\mathcal{E}}}\rvert - (i + k)$	$\lvert\boldsymbol{j}_{\widehat{\mathcal{E}}}\rvert - (i + k)$	$\lvert\boldsymbol{j}_{\widehat{\mathcal{E}}}\rvert$
$\lvert\boldsymbol{\mathcal{R}}_\omega^3\rvert$	0	0	0	0	$-(i + k)$

4.4. State sets and the transition matrix

In this section, we discuss the partitioning of the generator matrix into blocks based on the state space structure defined in Section 4.3 in order to model the network behaviors. Partitioning greatly reduces the modeling complexity of the transition behaviors by building matrix blocks having common transition patterns. We first define the primary set $H(n)$ to include all states that have n bed requests remaining in the network. Every state that satisfies the following equation can be categorized in the set $H(n)$:

$$|\boldsymbol{j}| + |\min(k,\ 0)| = n. \tag{2}$$

For example, both $\left(1,\ \left\langle \widehat{\mathcal{E}} \right\rangle,\ 0\right)$ and $\left(0,\ \langle\ \rangle,\ -1\right)$ are elements of the set $H(1)$ since there is a single bed request remaining in the network in each of these states (whether it is released or not). In state $\left(1,\ \left\langle \widehat{\mathcal{E}} \right\rangle,\ 0\right)$, there is an $E_\omega^{P\emptyset}$ patient for whom an advance bed request was sent, and preparation of the bed is in progress. In state $\left(0,\ \langle\ \rangle,\ -1\right)$, an $\mathcal{R}_\omega^3$ bed is waiting to be occupied.

To define the secondary sets, let $\Gamma_\nu(X)$ denote the set containing all permutations of elements of the set X with repetitions of size ν ($|\Gamma_\nu(X)| = |X|^\nu$). We partition each primary set into two secondary sets, $H(n)^-$ and $H(n)^+, \forall n \ge 1$, as follows:

1. $H(n)^- = \left\{(i, \boldsymbol{j}, k) :\ Eq.\ (1)\text{–}(2)\ \text{hold}, k < 0,\ \boldsymbol{j} \in \bigcup_{\nu=0}^{n} \Gamma_\nu(\{\widehat{\mathcal{E}}, \widetilde{\mathcal{E}}\})\right\}$, $\forall n \ge 1$.
2. $H(n)^+ = \left\{(i, \boldsymbol{j}, k)\ :\ Eq.\ (1)\text{–}(2)\ \text{hold},\ k \ge 0,\ \boldsymbol{j} \in \Gamma_n(\{\widehat{\mathcal{E}}, \widetilde{\mathcal{E}}\})\right\}$, $\forall n \ge 1$.

The state space is partitioned based on the sign of k because the network response to an activity depends on the type of entities at server s_ω^3, i.e., patients or beds. The secondary set $H(n)^-$ includes all states in which $\mathcal{E}_\omega^3$ or $\mathcal{R}_\omega^3$ beds wait for patients at server s_ω^3, while $H(n)^+$ represents the states in which E_ω patients wait for beds at server s_ω^3. Higher steady state probabilities for states in $H(n)^+$ imply that more patients are waiting for inpatient beds (due to prediction errors or shorter bed request lead times). In contrast, higher probabilities for states in $H(n)^-$ mean that more beds are prepared but remain unoccupied (due to prediction errors or longer bed request lead times.) The state space of the model is infinite in all three dimensions. We approximate the system by applying the truncation method (Green, 1984, 1985). This method captures the model's blocking probability based on the PASTA property (Wolff, 1982), with a truncation parameter ($\bar{\xi}$) set to determine how far the generator $\mathbf{G}$ should be expanded to reach a sufficient level of accuracy. The resulting generator $\mathbf{G}$ is based on the secondary sets as shown in Eq. (3) (see Box I), where the blank cells of the matrix are filled with $\mathbf{0}$. The generator $\mathbf{G}$ involves eleven unique types of blocks, examples of which can be seen in the $H(n)^-$ and $H(n)^+$ rows. For instance, states in $H(n)^-$ make transitions into states in $H(n-2)^+$ through $\mathbf{g}_{(n)^-}^{(n-2)^+}$. The size of each of the eleven blocks grows with a consistent pattern along all three dimensions (i, $\boldsymbol{j}$, and k) because keeping the signs in the subscript and superscript unchanged, $\mathbf{g}_{(n_1)^-}^{(n_2)^-}$ and $\mathbf{g}_{(n_1+\Delta)^-}^{(n_2+\Delta)^-}$ share an identical structure of transition patterns for any integer Δ. Since each block uniquely expands as n increases in accordance with the size of the system, $\mathbf{g}_{(n_1+\Delta)^-}^{(n_2+\Delta)^-}$ has extended patterns relative to $\mathbf{g}_{(n_1)^-}^{(n_2)^-}$ for $\Delta > 0$.

$$
\mathbf{G} = \begin{array}{c} \\ H(0) \\ \vdots \\ H(n-1)^- \\ H(n-1)^+ \\ H(n)^- \\ H(n)^+ \\ H(n+1)^- \\ H(n+1)^+ \\ \vdots \\ H(\bar{\xi})^+ \end{array}
\begin{array}{c}
\begin{array}{ccccccccccccc} H(0) & \cdots & H(n-2)^- & H(n-2)^+ & H(n-1)^- & H(n-1)^+ & H(n)^- & H(n)^+ & H(n+1)^- & H(n+1)^+ & \cdots & H(\bar{\xi})^+ \end{array} \\
\begin{bmatrix}
\mathbf{g}^{(0)^+}_{(0)^+} & \cdot & & & & & & & & & & \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & & & & & \\
 & \cdot & \mathbf{g}^{(n-2)^-}_{(n-1)^-} & \mathbf{g}^{(n-2)^+}_{(n-1)^-} & \mathbf{g}^{(n-1)^-}_{(n-1)^-} & \mathbf{g}^{(n-1)^+}_{(n-1)^-} & \mathbf{g}^{(n)^-}_{(n-1)^-} & & & & & \\
 & & & \mathbf{g}^{(n-2)^+}_{(n-1)^+} & \mathbf{g}^{(n-1)^-}_{(n-1)^+} & \mathbf{g}^{(n-1)^+}_{(n-1)^+} & \mathbf{0} & \mathbf{g}^{(n)^+}_{(n-1)^+} & & & & \\
 & & \mathbf{g}^{(n-2)^-}_{(n)^-} & \mathbf{g}^{(n-2)^+}_{(n)^-} & \mathbf{g}^{(n-1)^-}_{(n)^-} & \mathbf{g}^{(n-1)^+}_{(n)^-} & \mathbf{g}^{(n)^-}_{(n)^-} & \mathbf{g}^{(n)^+}_{(n)^-} & \mathbf{g}^{(n+1)^-}_{(n)^-} & & & \\
 & & & & & \mathbf{g}^{(n-1)^+}_{(n)^+} & \mathbf{g}^{(n)^-}_{(n)^+} & \mathbf{g}^{(n)^+}_{(n)^+} & \mathbf{0} & \mathbf{g}^{(n+1)^+}_{(n)^+} & & \\
 & & & & \mathbf{g}^{(n-1)^-}_{(n+1)^-} & \mathbf{g}^{(n-1)^+}_{(n+1)^-} & \mathbf{g}^{(n)^-}_{(n+1)^-} & \mathbf{g}^{(n)^+}_{(n+1)^-} & \mathbf{g}^{(n+1)^-}_{(n+1)^-} & \mathbf{g}^{(n+1)^+}_{(n+1)^-} & \cdot & \\
 & & & & & & & \mathbf{g}^{(n)^+}_{(n+1)^+} & \mathbf{g}^{(n+1)^-}_{(n+1)^+} & \mathbf{g}^{(n+1)^+}_{(n+1)^+} & \cdot & \\
 & & & & & & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 & & & & & & & & & & \cdot & \mathbf{g}^{(\bar{\xi})^+}_{(\bar{\xi})^+}
\end{bmatrix}
\end{array}. \tag{3}
$$

Box I.

Patient–bed assignment is affected by the types of beds that are present when any activity occurs within the ED-to-IU$_\omega$ network. Hence, we define tertiary sets based on the availability of the different types of beds to model the state space expansion patterns. For representation of the tertiary sets, without loss of generality, we focus on the $M/M/\infty - M/M/1$ network case in order to keep the expression concise and explicit; this case can be readily extended and applied to $M/M/\infty - M/M/c$ cases for any value of c. Depending on the availability of the different types of beds within the ED-to-IU$_\omega$ network, $H(n)^-$ can be categorized into n+2 tertiary sets as follows:

1. $H(n)^-_1 = \{(i, \boldsymbol{j}, k) \in H(n)^- : \ 0 \le i < n, \ \boldsymbol{j} = \emptyset, \ k = -n\}, \ \forall n \ge 1$. States in the tertiary set $H(n)^-_1$ have more beds than necessary for ED patients for whom bed requests have been sent, and the redundant reserved beds are released. There are no non-ED patients, and no more beds need to be prepared.
2. $H(n)^-_2 = \{(i, \boldsymbol{j}, k) \in H(n)^- : \ 0 \le i < n, \ \boldsymbol{j} = \langle \widehat{\mathcal{E}} \rangle, \ k = -n+1\}, \ \forall n \ge 2$. States in the tertiary set $H(n)^-_2$ have more beds than necessary for ED patients for whom bed requests have been sent, and the redundant reserved beds are released. There are no non-ED patients. The bed for which the preparation process has already started is released.
3. $H(n)^-_{3+\nu} = \{(i, \boldsymbol{j}, k) \in H(n)^- : i = |\boldsymbol{j}_{\widehat{\mathcal{E}}}| + |k|, \ \boldsymbol{j} \in \Gamma_\nu(\{\widehat{\mathcal{E}}, \ \widetilde{\mathcal{E}}\}), \ k = -n+\nu\}, \ \forall n \ge 2, \ 0 \le \nu \le n-1$. States in the tertiary set $H(n)^-_{3+\nu}$ have ν beds being prepared for $E^{P\emptyset}_\omega$, E^{PP}_ω, E^{NP}_ω, and $\widetilde{E}_\omega$ patients in the order in which the bed requests arrived. The number of patients for whom bed requests have been sent is equal to the sum of the numbers of beds that are being prepared and are ready. Hence, no bed is released.

The tertiary sets $H(n)^-_1, H(n)^-_2, \ \forall n \ge 1$, and $H(n)^-_{3+\nu}, \ \forall n \ge 2, \ 0 \le \nu \le n-1$, cover all states in the secondary set $H(n)^-$ for each n. Considering the permutations in $\boldsymbol{j}$ for $H(n)^-_{3+\nu}$, the number of states in the secondary set $H(n)^-$ can be calculated using the following formula:

$$X(n) = \sum_{j=1}^{n} \sum_{i=1}^{j} \frac{(j-1)!}{(j-i)!(i-1)!} + 2n = \sum_{j=1}^{n} 2^{j-1} + 2n, \quad \forall n \in \mathbb{Z}_+, \tag{4}$$

with a boundary condition $X(0) = 1$.

Each of the secondary sets $H(n)^+, \forall n \ge 1$, is classified into $|\Gamma_n(\{\widehat{\mathcal{E}}, \widetilde{\mathcal{E}}\})| = 2^n$ tertiary sets as follows:

$$H(n)^+_\tau = \{(i, \boldsymbol{j}, k) \in H(n)^+ : i = |\boldsymbol{j}_{\widehat{\mathcal{E}}}| - k, \ \boldsymbol{j} = \Gamma_n(\{\widehat{\mathcal{E}}, \widetilde{\mathcal{E}}\})_\tau, \ 0 \le k \le |\boldsymbol{j}_{\widehat{\mathcal{E}}}|\}, \tau = 1, \ldots, 2^n, \forall n \ge 1,$$

where $\Gamma_n(\{\widehat{\mathcal{E}}, \widetilde{\mathcal{E}}\})_\tau$ is the τth element of the colexicographically ordered set $\Gamma_n(\{\widehat{\mathcal{E}}, \widetilde{\mathcal{E}}\})$, with $\widehat{\mathcal{E}} < \widetilde{\mathcal{E}}$. For example, for $n = 2$, we have the colexicographically ordered $\Gamma_2(\{\widehat{\mathcal{E}}, \widetilde{\mathcal{E}}\}) = \{\langle \widehat{\mathcal{E}}\widehat{\mathcal{E}} \rangle, \langle \widetilde{\mathcal{E}}\widehat{\mathcal{E}} \rangle, \langle \widehat{\mathcal{E}}\widetilde{\mathcal{E}} \rangle, \langle \widetilde{\mathcal{E}}\widetilde{\mathcal{E}} \rangle\}$ and $\Gamma_2(\{\widehat{\mathcal{E}}, \widetilde{\mathcal{E}}\})_3 = \langle \widehat{\mathcal{E}}\widetilde{\mathcal{E}} \rangle$. Table 5 summarizes the bed availability information utilized in partitioning the tertiary sets.

Based on the structure of the tertiary set $H(n)^+_\tau$, the number of states in the secondary set $H(n)^+$ is calculated as follows:

$$Y(n) = \sum_{j=1}^{n+1} \sum_{i=1}^{j} \frac{n!}{(n-i+1)!(i-1)!}, \quad \forall n \in \mathbb{Z}_{\ge 0}. \tag{5}$$

The sum of $X(n)$ and $Y(n)$, denoted as $Z(n)$, gives the number of states in the primary set $H(n)$. The total number of states in the network that allow N bed requests is calculated in Eq. (6):

$$\boldsymbol{\Phi}(N) = \sum_{n=0}^{N} \big[X(n) + Y(n)\big], \quad \forall n \in \mathbb{Z}_{\ge 0}. \tag{6}$$

The generator $\mathbf{G}$ is the fully specified ordering method based on bed availability that is described in this section. More details about the structure of each $\mathbf{g}$ block are presented in Online Appendices OA.2 and OA.3. The model is solved by determining the truncation parameter $\bar{\xi}$. Let $\mathrm{E}[L^P_\omega(\xi)]$ be the mean number of patients in server s^3_ω with at most ξ bed requests remaining within the network; $\mathrm{E}[L^B_\omega(\xi)]$ is similarly defined for the beds in server s^3_ω. We define two parameters as follows:

Table 5
Bed Availability Information for States in Each Tertiary Set.

	$\mathcal{E}^2_\omega$	$\mathcal{R}^2_\omega$	$\widetilde{\mathcal{E}}^2_\omega$	$\mathcal{E}^3_\omega$	$\mathcal{R}^3_\omega$	$\lvert\mathcal{E}^2_\omega\rvert + \lvert\mathcal{R}^2_\omega\rvert + \lvert\widetilde{\mathcal{E}}^2_\omega\rvert$	$\lvert\mathcal{E}^3_\omega\rvert + \lvert\mathcal{R}^3_\omega\rvert$
$H(n)^-$							
$H(n)^-_1$	×	×	×	○	○		
$H(n)^-_2$	×	○	×	○	○	$n-k$	k
$H(n)^-_{3+\nu}$	○	×	○	○	×		
$H(n)^+$							
$H(n)^+_\tau$	○	○	○	×	×	n	0

○: available, ×: unavailable.

$$\begin{aligned}
&\mathrm{E}[L_\omega^P(\xi)] = \sum_{n=1}^{\xi} \sum_{s=\Phi(n-1)+X(n)+1}^{\Phi(n)} k_{(\theta)}\pi_{(\theta)} + \sum_{s=1}^{\Phi(\xi)} |\boldsymbol{j}_{\widetilde{\mathcal{E}}(\theta)}|\pi_{(\theta)},\\
&\mathrm{E}[L_\omega^B(\xi)] = \sum_{n=1}^{\xi} \sum_{s=\Phi(n-1)+1}^{\Phi(n-1)+X(n)} (-k_{(\theta)})\pi_{(\theta)},
\end{aligned} \tag{7}$$

to obtain

$$\begin{aligned}
\xi^P &= \min_{\xi}\left\{\xi : \left|\mathrm{E}[L_\omega^P(\xi)] - \mathrm{E}[L_\omega^P(\xi-1)]\right| / \mathrm{E}[L_\omega^P(\xi)] < \epsilon\right\},\\
\xi^B &= \min_{\xi}\left\{\xi : \left|\mathrm{E}[L_\omega^B(\xi)] - \mathrm{E}[L_\omega^B(\xi-1)]\right| / \mathrm{E}[L_\omega^B(\xi)] < \epsilon\right\},\\
\overline{\xi} &= \max\{\xi^P, \xi^B\},
\end{aligned} \tag{8}$$

where $\pi_{(\theta)}$ is the stationary probability of state θ, and $k_{(\theta)}$ and $\boldsymbol{j}_{\widetilde{\mathcal{E}}(\theta)}$ are the values of k and $\boldsymbol{j}_{\widetilde{\mathcal{E}}}$ for state θ, respectively. ϵ is the truncation criterion, which is set to obtain a specified level of numerical stability in the mean numbers of patients and beds in server s_ω^3. For our experiments, we set $\epsilon = 0.02$ to keep the approximation errors at a negligible level. Algorithm 1 derives the truncation parameter and constructs the corresponding generator matrix. To check on the detailed structure of the generator matrix $\mathbf{G}_{\Phi(m)\times\Phi(m)}$, refer to Online Appendices OA.2 and OA.3. The algorithm and systems of equations constructed in the form of matrix (as shown in Online Appendix OA.3) are built in the R programming language.

Algorithm 1 Constructing and Solving Generator Matrix **G**

1: Set ϵ
2: $m = 0$
3: **while** $\left|\mathrm{E}[L_\omega^P(m)] - \mathrm{E}[L_\omega^P(m-1)]\right| / \mathrm{E}[L_\omega^P(m)] > \epsilon$ and $\left|\mathrm{E}[L_\omega^B(m)] - \mathrm{E}[L_\omega^B(m-1)]\right| / \mathrm{E}[L_\omega^B(m)] > \epsilon$ **do**
4: $m \leftarrow m + 1$
5: Create $\mathbf{G}_{\Phi(m)\times\Phi(m)} = \mathbf{0}$
6: **for** integer n in $[0, m-1]$ **do**
7: Construct all the $\mathbf{g}$ matrices for n and fill the values in $\mathbf{G}_{\Phi(m)\times\Phi(m)}$
8: **end for**
9: Construct a subset of the $\mathbf{g}$ matrices for m (excluding $\mathbf{g}_{(m)}^{(m+\Delta)}$ for $\Delta > 0$) and fill the values in $\mathbf{G}_{\Phi(m)\times\Phi(m)}$
10: Obtain the vector $\boldsymbol{\pi}$ by solving the equations $\boldsymbol{\pi}\mathbf{G}_{\Phi(m)\times\Phi(m)} = \mathbf{0}$, $\boldsymbol{\pi}\mathbf{e} = 1$
11: Obtain $\mathrm{E}[L_\omega^P(m)]$ and $\mathrm{E}[L_\omega^B(m)]$ by solving Eq. (8) using notation from Eq. (7)
12: **end while**
13: Calculate to obtain $\mathrm{E}(L_\omega^E)$, $\mathrm{E}(L_\omega^{\widetilde{E}})$, $\mathrm{E}(L_\omega^B)$, $\mathrm{E}(W_\omega^E)$, $\mathrm{E}(W_\omega^{\widetilde{E}})$

4.5. The multiple inpatient unit case

A general hospital consists of multiple IUs, and admissions from the ED to these multiple IUs are interdependent with respect to the patient flow. First, the sum of the admission rates to individual IUs equals the total rate of admissions to the hospital. Moreover, an E_ω^{PN} patient in IU ω is admitted to another IU, say IU ω', as an $E_{\omega'}^{NP}$ patient. Let Ω be a set containing all IUs for bed allocation. The following equations express relationships between the different IUs and are used to model the bed allocation operations across multiple IUs:

$$\begin{aligned}
&\text{(a)} \quad \sum_{\omega\in\Omega} p_\omega = \sum_{\omega\in\Omega} Z_\omega = 1,\\
&\text{(b)} \quad \sum_{\omega\in\Omega} p_\omega(1-q_\omega) = \sum_{\omega\in\Omega} r_\omega(1-p_\omega).
\end{aligned} \tag{9}$$

Members of the set Ω are mutually exclusive and collectively exhaustive with respect to disposition decision predictions as well as the true disposition decisions. Therefore, Eq. (9)(a) holds. Eq. (9)(b) indicates that the total amounts of Type-I and Type-II errors over the set Ω should always be balanced. Together, Eqs. (9)(a)–(b) formulate the relationships among the parameters p_ω, q_ω, r_ω, and Z_ω over the set Ω for modeling the ED-to-IU$_\Omega$ network.

Meanwhile, even though there are multiple IUs through which patients move according to their prediction results (conforming to Eqs. (9)(a)–(b)), the patient flow over the entire ED-to-IU$_\Omega$ network is still a feed-forward flow since patients do not revisit the same server or queue during their admission processes. Therefore, according to Burke's output theorem (Burke, 1956), the steady-state distribution of the system can be evaluated based on the property of the independence of the queue lengths as given by Eq. (10):

$$\pi(\omega_1 = L_1^{(\cdot)}, \omega_2 = L_2^{(\cdot)}, \ldots, \omega_\delta = L_\delta^{(\cdot)}) = \prod_{i=1}^{\delta} \pi(\omega_i = L_i^{(\cdot)}), \tag{10}$$

where we assume a total of δ dispositions and $L_i^{(\cdot)}$ represents the queue length for entity $(\cdot)$ at ω_i. Relying on these relationships, the analysis results for the multiple IU groups are produced independently and presented in Section 5.4.

5. Analysis of the model

In this section, we evaluate the effectiveness of our proposed proactive approach to bed allocation. Our primary performance measure is the inpatient bed allocation wait time for both types of patients. We analyze the system under two main scenarios: simple and complex. Under the simple scenario, we assume that all beds in IU ω are dedicated to E_ω patients and that the predictions for these patients are perfect. The purpose of evaluating the simple case is to explicitly and concisely show the fundamental impact of the proposed strategy on reducing the inpatient bed allocation delays for ED patients. In contrast, under the complex scenario, we assume that IU ω serves both E_ω and $\widetilde{E}_\omega$ patients and that the predictions for ED patients are error prone. We also examine the trade-off relationship between proactivity and prediction quality by analyzing the models with the prediction performance measures obtained at real-world settings.

5.1. Simple scenario: IU dedicated to ED patients with perfect prediction

For the simple scenario, we assume that λ^E is 10 patients/hour, and admission rate for IU ω, i.e., $\lambda^E Z_\omega$, is 0.4 patients/hour, mimicking an IU group managed by an EVS server at the study hospital. The expected bed allocation delays for various bed request signal lead times ($1/\mu_\omega^1$) and bed preparation service times ($1/\mu_\omega^2$) are displayed in Fig. 8. We observe that when the proactive bed allocation scheme is implemented

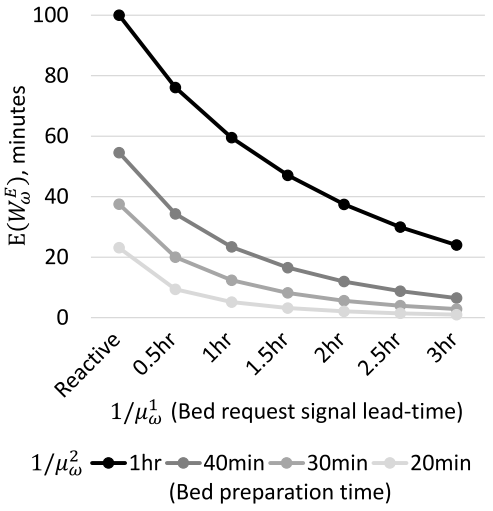


Fig. 8. Impact of advance bed requests for E_ω patients admitted to ED-dedicated IU$_\omega$ with perfect prediction.

Table 6
Impact of proactive bed allocation for E_ω and $\widetilde{E}_\omega$ patients with perfect disposition prediction.

(a) $E(W_\omega^E)$

$1/\mu_\omega^1$ \ u_ω	1	0.75	0.5	0.25	0
Reactive	100.00	100.00	100.00	100.00	-
0.5hr	76.08	76.25	76.44	76.64	-
1hr	59.48	60.09	60.77	61.55	-
1.5hr	47.03	48.11	49.36	50.83	-
2hr	37.43	38.89	40.65	42.78	-
2.5hr	29.91	31.66	33.82	36.52	-
3hr	23.97	25.90	28.35	31.51	-

(b) $E(W_\omega^{\widetilde{E}})$

$1/\mu_\omega^1$ \ u_ω	1	0.75	0.5	0.25	0
Reactive	-	100.00	100.00	100.00	100.00
0.5hr	-	99.93	99.93	99.93	99.93
1hr	-	99.93	99.93	99.93	99.93
1.5hr	-	99.93	99.93	99.93	99.93
2hr	-	99.93	99.93	99.93	99.93
2.5hr	-	99.93	99.93	99.93	99.93
3hr	-	99.93	99.93	99.93	99.93

Unit: minutes

with a $1/\mu_\omega^1$ of 1.5 h for the IU case in which $1/\mu_\omega^2$ is 1 h, the bed allocation delays can be reduced from 100 min (under the reactive strategy) to 47 min, i.e., there is a 53% reduction. The figure also shows that an IU with 40 min of $1/\mu_\omega^2$ guided by advance bed request signals with a $1/\mu_\omega^1$ of 1 h experiences the same bed allocation delays as an IU operating with only 20 min of $1/\mu_\omega^2$ without receiving advance bed request signals. Moreover, Fig. 8 shows that applying proactive coordination to a busier ED-to-IU network leads to a larger reduction (in terms of length) in bed allocation delays.

5.2. Complex scenario (1): E_ω and $\widetilde{E}_\omega$ patient admissions with perfect prediction

Our framework aims to retain the level of the bed allocation delays for $\widetilde{E}_\omega$ patients while proactively reserving inpatient beds for E_ω patients. To explicitly represent this, we evaluate the case where there are no prediction errors, i.e., $\mathcal{R}_\omega^2$ and $\mathcal{R}_\omega^3$ beds never appear within the ED-to-IU$_\omega$ network. While $1/\mu_\omega^2$ is set 1 h, the impact of u_ω (the proportion of admissions from the ED among all admissions) on $E(W_\omega^E)$ is shown in Table 6(a). It can be seen that there are additional gains in reducing the bed allocation delays for E_ω patients as u_ω becomes larger. This is because when there are more E_ω patients (with a larger u_ω), the advantage of reserving IU beds for the group of E_ω patients is further enhanced. For the same reason, the chance that an E_ω patient will benefit from occupying an IU bed that is reserved for an E_ω patient becomes even greater as $1/\mu_\omega^1$ increases (at a higher u_ω level). This effect is not present for patients for whom proactive bed preparation and reservations are not practiced (the $\widetilde{E}_\omega$ patients in Table 6(b)). The numerical results derived by our model also confirm that there is no change in $E(W_\omega^{\widetilde{E}})$ regardless of the values of μ_ω^1 and u_ω under the perfect prediction setting, by the property of Poisson processes (Table 6(b)). The negligible difference between the theoretical and numerical results (4 s) comes from the truncation-based approximation discussed in Section 4.4.

5.3. Complex scenario (2): E_ω and $\widetilde{E}_\omega$ patient admissions with imperfect prediction

To evaluate the effect of erroneous predictions on the operational performance measures, we conduct analyses under the setting where $1/\mu_\omega^2$ is 1 h and u_ω is 0.5. We assume that the frequency of advance bed requests is the same as the actual admission rate from the ED to unit ω, i.e., $p_\omega = Z_\omega$, even though individual predictions could be wrong. Figs. 9(a) and (b) display $E(W_\omega^E)$ and $E(W_\omega^{\widetilde{E}})$, respectively, under the various positive predictive values (σ_ω^{PPV}) in Fig. 5. We note that when $p_\omega = Z_\omega$, $\sigma_\omega^{PPV} = \sigma_\omega^{TPR}$. As the quality of predictions decreases as σ_ω^{PPV} and σ_ω^{TPR} become smaller, the impact of proactive bed allocation on $E(W_\omega^E)$ decreases (represented by the different line graphs in Fig. 9(a)). The cost of prediction errors incurred for E_ω patients can be approximated by the gap between the lines for $\sigma_\omega^{PPV} < 1$ and $\sigma_\omega^{PPV} = 1$ at each μ_ω^1 level. On the other hand, the released beds can be taken by $\widetilde{E}_\omega$ patients with no delay, resulting in an unintended reduction of bed allocation delays for $\widetilde{E}_\omega$ patients (Fig. 9(b)). Hence, bed allocation delays for $\widetilde{E}_\omega$ patients under the proposed proactive bed allocation approach are bounded by the delays in the reactive case. Moreover, the delays for $\widetilde{E}_\omega$ patients are reduced as σ_ω^{PPV} and σ_ω^{TPR} decrease (due to an increased number of released beds generated by false positives). However, the level of delay reductions quickly saturates as $1/\mu_\omega^1$ increases, without extended benefits. In other words, the chance of $\widetilde{E}_\omega$ patients taking $\mathcal{R}_\omega^2$ or $\mathcal{R}_\omega^3$ beds does not increase with higher $1/\mu_\omega^1$ values. This is because our proposed bed reservation rules restrict the number of bed requests in the network based on the actual bed demand and do not allow an excessive number of redundant bed requests to ensure efficient EVS utilization. We also present $E(L_\omega^B)$—the expected number of beds that are ready to be occupied in server s_ω^3—in Fig. 9(c) as evidence of the inefficiency caused by prediction errors. While having available beds in server s_ω^3 is not an issue in most cases and is instead recommended, having an excessive number of

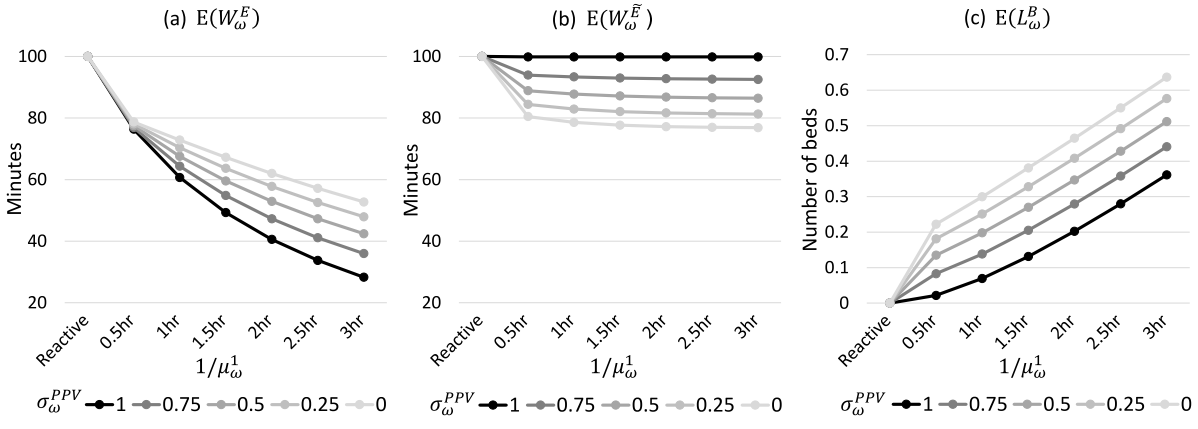


Fig. 9. Impact of proactive bed allocation on wait times of patients and beds with imperfect prediction.

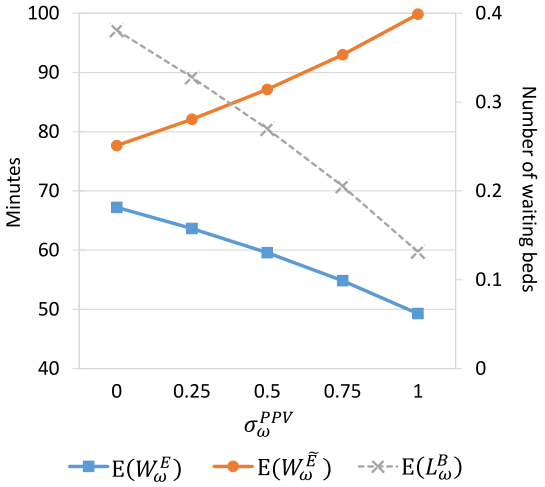


Fig. 10. Summary of operational performance measures depending on prediction quality.

unoccupied beds during the high utilization of EVS staff would indicate inefficient EVS operations owing to false positive bed request signals.

Fig. 10 comprehensively represents the impact of the quality of predictions on the three performance measures. We select the setting where $1/\mu_\omega^1 = 1.5$ hour, $1/\mu_\omega^2 = 1$ hour, $u_\omega = 0.5$, and $p_\omega = Z_\omega$. For $\sigma_\omega^{PPV} = \sigma_\omega^{TPR} < 1$, E_ω and $\widetilde{E}_\omega$ patients compete for $\mathcal{R}_\omega^2$ and $\mathcal{R}_\omega^3$ beds generated by false positives, and released beds are shared between the two types of patients based on the value of u_ω as well as the quality of the predictions. While $\widetilde{E}_\omega$ patients could benefit from false positives, $E(L_\omega^B)$ increases as the prediction quality decreases.

5.4. Trade-off between extent of proactivity and prediction quality with progressive predictions

Finally, we present the most realistic cases (inspired by analyzing patient information from the case study hospital) to see how the bed allocation delays for E_ω patients can vary depending on the quality of the predictions, which evolve progressively throughout the ED caregiving cycle. In particular, we identify four essential discrete caregiving epochs in the ED: triage, initial assessment, return of initial laboratory test results, and the disposition decision, with a significant amount of information gathered at each epoch. While additional information can be revealed any time, these four epochs are fairly well-established and are common for most ED patients in most hospitals. As a patient undergoes more ED caregiving processes, the amount of clinical information grows and can improve the reliability of the prediction, thus elevating prediction performance measures. However, a delayed/postponed bed request signal can compromise the timeliness of the bed allocation processes that follow. Table 7 reveals how this trade-off can affect the reduction of bed allocation delays in different ways by comparing analysis results from two distinct IUs ((a) the TU and (b) the ICU) that have different trajectories for prediction quality progress.

The caregiving epochs, the average remaining LoS (i.e., $1/\mu_\omega^1$) during each epoch, and the performance measures for disposition decision predictions (σ_ω^{PPV}, σ_ω^{TPR}, and σ_ω^{FPR}) for each unit are displayed in Table 7. Since there is no significant difference in the values of $1/\mu_\omega^1$ for E_{TU} and E_{ICU} patients in each epoch, we assume that $1/\mu_\omega^1$ during each epoch is identical for E_{TU} and E_{ICU} patients in order to make a concise comparison. We built multinomial logistic regression models, exploiting the EHR data from the study hospital for 172,809 patients between May 2014 and April 2016. We used the first 85% of the dataset for training the classification models and the rest for testing. The performance measures were derived by fitting the testing sample. Unlike the previous analyses, the p_ω values obtained from the actual prediction results turn out to be strictly less than Z_ω for all of the epochs, i.e., $p_\omega/Z_\omega < 1$. All of the other parameters remain the same as in Section 5.3, to avoid confusion.

E_{TU} patients seem to have less clinical information that can be used to classify them during triage (vital signs, complaints, and so on) compared to E_{ICU} patients (Table 7(a)). However, waiting for additional clinical information from the downstream processes leads to considerably improved predictive capabilities for E_{TU} patients. As a result, making bed request decisions after receiving initial laboratory test results can further reduce $E(W_{TU}^E)$ to 52 min, compensating above and beyond for the negative influence of postponement. In contrast, as shown in Table 7(b), more clinically distinctive information is gathered for E_{ICU} patients at triage, which generates comparatively reliable prediction results within about 20 min of E_{ICU} patient arrival at the ED, on average. Relatively speaking, the most significant information for E_{ICU} patients is already revealed at triage, and the information gained afterward does not drastically improve prediction quality for the patient group. As a result, there is no operational benefit in postponing bed request signals in order to, for instance, acquire laboratory test results for improving predictions for E_{ICU} patients. The list of data items gathered during each caregiving epoch is presented in more detail in Online Appendix OA.4.

6. Discussion and conclusion

To remedy the growing overcrowding that is being witnessed in EDs and the negative consequences associated with it, we propose an ETI scheme that facilitates proactive inpatient bed allocations for ED

Table 7
Evolution of prediction quality throughout ED processes and its operational impact.

(a) E_{TU} patients	Triage	Initial assessment	Initial lab results	Disposition decision
$1/\mu_{TU}^1$	250 min	200 min	145 min	0 min
p_{TU}/Z_{TU}	0.50	0.78	0.93	1.00
σ_{TU}^{PPV}	0.24	0.31	0.42	1.00
σ_{TU}^{TPR}	0.12	0.24	0.39	1.00
σ_{TU}^{FPR}	0.016	0.022	0.022	1.00
$E(W_{TU}^E)$	63 min	53 min	**52 min**	100 min

(b) E_{ICU} patients	Triage	Initial assessment	Initial lab results	Disposition decision
$1/\mu_{ICU}^1$	250 min	200 min	145 min	0 min
p_{ICU}/Z_{ICU}	0.81	0.82	0.84	1.00
σ_{ICU}^{PPV}	0.49	0.52	0.56	1.00
σ_{ICU}^{TPR}	0.40	0.43	0.47	1.00
σ_{ICU}^{FPR}	0.017	0.016	0.015	1.00
$E(W_{ICU}^E)$	**42 min**	47 min	53 min	100 min

patients by utilizing prediction information. The proposed bed preparation and allocation rules can be used to develop an automated proactive bed management system that judiciously utilizes ED disposition decision predictions and LoS estimation information in hospitals. Indeed, the results from our study inspired the study hospital and collaborators to implement the proposed ETI strategy within its ED-to-IU network. A data processing/analytics platform is currently under development to enable proactive inpatient bed preparation and allocation at the hospital.

Our proposed proactive bed allocation approach can be further studied for operationalization. First, as can be seen from Fig. 1, the ED-to-IU network is a dynamic system with varying patient arrival and discharge rates, possibly requiring non-stationary analysis. While a simulation study is recommended based on the complexity of the network and the detailed bed management strategy proposed in this paper, we focused on incorporating the key aspect of the network and inpatient bed allocation rules and generating original insights by evaluating classification performance in terms of operational benefits. For all practical purposes, a beginning strategy would be to assume that the non-stationary rates can be effectively characterized as a relatively small number of distinct steady-state systems (each spanning one to several hours). Once the network is so characterized, it can implement the advance inpatient bed request signals based on the findings of this study, further calibrate the parameters, and update the optimal policy based on its own operational situation.

Second, the proposed coordination scheme can inspire different types of process improvements in the ED-to-IU network. For instance, when EVS servers are shared as a completely pooled server system across all IUs, deployment of the servers becomes a prioritization problem. In this setting, a hospital can utilize proactive bed allocation schemes to strategically deploy its servers according to predicted inpatient bed demand. Since reactive processes are prevalent in current ED-to-IU network operations across most hospitals (e.g., admission approval, administrative procedures, and transporter assignment), proactive coordination methods can contribute in different forms and ways depending on the specific ED-to-IU operations and practices of different hospitals. Although different settings exist, this paper provides core ideas, a rigorous representation, and an operational impact analysis of proactive bed allocation in a large hospital setting.

Finally, even though this paper focuses on a healthcare operations setting, it adds to the general body of literature investigating service systems that can benefit from proactive resource coordination based on prediction outcomes (e.g., just-in-time logistics, manufacturing, and project management). The implementation of ETI in complex service systems should become a promising area of scientific research and exploration for both industry and academia.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijpe.2020.107842. This manuscript has an online appendix that offers additional technical discussions of the modeling ideas and methodology for attaining analytical tractability for real-world ED-to-IU network settings as well as information about the data items used to build the multinomial logistic regression models for disposition decision prediction.

References

Abelson, R., 2013. E.R.'s Account for Half of Hospital Admissions, Study Says. The New York Times, New York, NY, http://www.nytimes.com/2013/05/21/business/half-of-hospital-admissions-from-emergency-rooms.html. (Accessed: 2019-06-10).

Araz, O.M., Olson, D., Ramirez-Nafarrate, A., 2019. Predictive analytics for hospital admissions from the emergency department using triage information. Int. J. Prod. Econ. 208, 199–207.

Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y.N., Tseytlin, Y., Yom-Tov, G.B., 2015. On patient flow in hospitals: A data-based queueing-science perspective. Stoch. Syst. 5 (1), 146–194.

Augustine, J.J., 2016a. 2015 emergency department survey shows spike in volume, structural changes, patient boarding concerns. http://www.acepnow.com/article/2015-emergency-department-survey-shows-spike-volume-structural-changes-patient-boarding-concerns/. (Accessed: 2019-06-10).

Augustine, J.J., 2016b. Long emergency department boarding times drive walkaways, revenue losses. http://www.acepnow.com/article/now-boarding/. (Accessed: 2019-06-10).

Batt, R.J., Terwiesch, C., 2017. Early task initiation and other load-adaptive mechanisms in the emergency department. Manag. Sci. 63 (11), 3531–3551.

Burke, P.J., 1956. The output of a queuing system. Oper. Res. 4 (6), 699–704.

Carter, E.J., Pouch, S.M., Larson, E.L., 2014. The relationship between emergency department crowding and patient outcomes: A systematic review. J. Nurs. Scholarsh. 46 (2), 106–115.

Casalino, E., Wargon, M., Peroziello, A., Choquet, C., Leroy, C., Beaune, S., Pereira, L., Bernard, J., Buzzi, J.C., 2014. Predictive factors for longer length of stay in an emergency department: a prospective multicentre study evaluating the impact of age, patient's clinical acuity and complexity, and care pathways. Emerg. Med. J. 31 (5), 361–368.

Chaou, C.H., Chen, H.H., Chang, S.H., Tang, P., Pan, S.L., Yen, A.M.F., Chiu, T.F., 2017. Predicting length of stay among patients discharged from the ED using an accelerated failure time model. PLoS One 12 (1), p.e0165756.

CMS, 2018. Hospital compare datasets, https://data.medicare.gov/data/hospital-compare/. (Accessed 11 June 2018).

CMS and Joint Commission, 2017. Specifications manual for national hospital inpatient quality measures. https://www.jointcommission.org/specifications_manual_for_national_hospital_inpatient_quality_measures.aspx. (Accessed: 2018-08-17).

Fatovich, D.M., Nagree, Y., Sprivulis, P., 2005. Access block causes emergency department overcrowding and ambulance diversion in Perth. West. Aust. Emerg. Med. J. 22 (5), 351–354.

George, F., Evridiki, K., 2015. The effect of emergency department crowding on patient outcomes. Health Sci. J. 9 (1), 1–6.

Golmohammadi, D., 2016. Predicting hospital admissions to reduce emergency department boarding. Int. J. Prod. Econ. 182, 535–544.

Green, L., 1984. A queueing system with auxiliary servers. Manag. Sci. 30 (10), 1207–1216.

Green, L., 1985. A queueing system with general-use and limited-use servers. Oper. Res. 33 (1), 168–182.

Henry, J., Pylypchuk, Y., Searcy, T., Patel, V., 2016. Adoption of EHR Systems Among U.S. Non-Federal Acute Care Hospitals: 2008-2015. The Office of National Coordination for Health Information Technology, (ONC/HIT Data Brief), https://www.healthit.gov/sites/default/files/briefs/2015_hospital_adoption_db_v17.pdf. (Accessed: 2018-06-10).

Hoot, N.R., Aronsky, D., 2008. Systematic review of emergency department crowding: Causes, effects, and solutions. Ann. Emerg. Med. 52 (2), 126–136.

Jelinek, G.A., Weiland, .T.J., Mackinlay, C., 2010. Supervision and feedback for junior medical staff in Australian emergency departments: Findings from the emergency medicine capacity assessment study. BMC Med. Educ. 10 (1), 74.

Kellermann, A.L., Jones, S.S., 2013. What it will take to achieve the as-yet-unfulfilled promises of health information technology. Health Aff. 32 (1), 63–68.

Kocher, K.E., Meurer, W.J., Desmond, J.S., Nallamothu, B.K., 2012. Effect of testing and treatment on emergency department length of stay using a national database. Acad. Emerg. Med. 19 (5), 525–534.

Lee, S.-Y., Chinnam, R.B., Dalkiran, E., Krupp, S., Nauss, M., 2019. Prediction of emergency department patient disposition decision for proactive resource allocation for admission. Health Care Manag. Sci. 23, 1–21.

Liu, S., Hobgood, C., Brice, J.H., 2003. Impact of critical bed status on emergency department patient flow and overcrowding. Acad. Emerg. Med. 10 (4), 382–385.

Peck, J.S., Benneyan, J.C., Nightingale, D.J., Gaehde, S.A., 2012. Predicting emergency department inpatient admissions to improve same-day patient flow. Acad. Emerg. Med. 19 (9), 1045–1054.

Peck, J.S., Benneyan, J.C., Nightingale, D.J., Gaehde, S.A., 2014. Characterizing the value of predictive analytics in facilitating hospital patient flow. IIE Trans. Healthc. Syst. Eng. 4 (3), 135–143.

Pines, J.M., Batt, R.J., Hilton, J.A., Terwiesch, C., 2011. The financial consequences of lost demand and reducing boarding in hospital emergency departments. Ann. Emerg. Med. 58 (4), 331–340.

Qiu, S., Chinnam, R.B., Murat, A., Batarse, B., Neemuchwala, H., Jordan, W., 2015. A cost sensitive inpatient bed reservation approach to reduce ED boarding times. Health Care Manag. Sci. 18 (1), 67–85.

Saghafian, S., Austin, G., Traub, S.J., 2015. Operations research/management contributions to ED patient flow optimization: Review and research prospects. IIE Trans. Healthc. Syst. Eng. 5 (2), 101–123.

Shi, P., Chou, M.C., Dai, J.G., Ding, D., Sim, J., 2015. Models and insights for hospital inpatient operations: Time-dependent ED boarding time. Manag. Sci. 62 (1), 1–28.

Song, H., Tucker, A.L., Murrell, K.L., 2015. The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. Manag. Sci. 61 (12), 3032–3053.

Thomas, B.G., Bollapragada, S., Akbay, K., Toledano, D., Katlic, P., Dulgeroglu, O., Yang, D., 2013. Automated bed assignments in a complex and dynamic hospital environment. Interfaces 43 (5), 435–448.

U.S. GAO, 2003. Hospital Emergency Crowded Conditions Vary Among Hospitals and Communities. U.S. General Accounting Office, http://www.gao.gov/new.items/d03460.pdf. (Accessed: 2018-06-10).

Van der Vaart, T., Vastag, G., Wijngaard, J., 2011. Facets of operational performance in an emergency room (ER). Int. J. Prod. Econ. 133 (1), 201–211.

Wertheimer, B., Jacobs, R.E., Bailey, M., Holstein, S., Chatfield, S., Ohta, B., Horrocks, A., Hochman, K., 2014. Discharge before noon: An achievable hospital goal. J. Hosp. Med. 9 (4), 210–214.

Wolff, R.W., 1982. Poisson arrivals see time averages. Oper. Res. 30 (2), 223–231.

Zhou, J.C., Pan, K.H., Zhou, D.Y., Zheng, S.W., Zhu, J.Q., Xu, Q.P., Wang, C.L., 2012. High hospital occupancy is associated with increased risk for patients boarding in the emergency department. Am. J. Med. 125 (4), 416.e1-7.