When to Use Provider Triage in Emergency Departments

Michael F. Kamali
University of Rochester Medical Center, Rochester, New York 14642
michael_kamali@urmc.rochester.edu

Tolga Tezcan
London Business School, Regent's Park, London NW1 4SA, UK
ttezcan@london.edu

Ozlem Yildiz
Darden School of Business, University of Virginia, Charlottesville, Virginia 22903
yildizo@darden.virginia.edu

We study triage decisions in emergency departments (EDs) and provide a general procedure for determining when to apply provider triage (PT) based on operational and financial considerations using a steady-state many-server fluid approximation. We then apply the proposed method in the setting of a teaching hospital's ED and obtain closed-form expressions for the range of arrival rates for which PT outperforms the traditional nurse triage economically. We show that the proposed solution methodology based on this approximation procedure is asymptotically optimal under a many-server asymptotic regime. We also demonstrate via simulation experiments that the proposed policy performs within 0.82% of the best solution obtained via a computationally intensive total enumeration method.

1. Introduction

Emergency department (ED) crowding has become a significant obstacle to providing timely emergency care in the U.S. in the last decade due to steadily increasing ED visits per year (GAO 2009, Pitts et al. 2012). ED crowding contributes to increased waiting times, patient dissatisfaction, ambulance diversion, higher rates of medical errors, increased mortality, and more patients leaving the EDs without being seen (see Pines et al. 2008, Pitts et al. 2012, Batt and Terwiesch 2015). The U.S. Government Accountability Office (GAO) reported that the waiting time for patients in EDs in 2006 exceeded the recommended time frame in 50.4% of cases (GAO 2009). In this paper, our goal is to analyze *provider triage (PT)* method, which is one of the interventions designed to alleviate the problems arising from ED crowding by reducing throughput time.

The process in the ED consists of two main stages: triage and treatment. Traditionally, triage is conducted by one or more nurses, who are referred to as triage nurses, in a method known as *nurse triage (NT)*. When patients arrive in the ED, they are usually triaged within a few minutes by a

triage nurse, who interviews the patient and records her medical history and complaints. Based on the information obtained, the triage nurse assigns the patient an acuity level and orders basic diagnostic tests, such as electrocardiograms (EKGs), if needed. After the triage is completed, the patient waits in the waiting room until a bed in the treatment area becomes available. Once the patient is assigned a bed, she is taken to the treatment area. There, she is evaluated for the first time by a provider, such as a physician, physician assistant, or nurse practitioner, who will administer the treatment. Once treated, the patient is either discharged or transferred to an inpatient unit, potentially after waiting some time in the ED. Although some patients, such as psychiatric or trauma patients, may follow a slightly different route, most patients follow the steps described above.

An alternative triage method is PT, which is also referred to as physician triage or team triage when triage is conducted by a team including a provider (see Saghafian et al. (2015) for a review of triage interventions). In PT, a *triage provider* performs triage in addition to the triage nurse. The triage provider performs a brief initial assessment or medical screening examination and initiates diagnostic testing and treatment in the triage area when necessary (Wiler et al. 2010). Thus, when PT is used, the patient is seen by a provider for the first time at the triage stage instead of the treatment stage as in NT. After PT, patients with only minor complaints (nearly 30% of patients who arrive in the ED, see Cooke et al. 2003) can be discharged after the initial evaluation in the triage area (Subash et al. 2004, Terris et al. 2004, Choi et al. 2006, Travers and Lee 2006). On the other hand, once their triage interventions are completed, patients with more severe conditions are sent to a waiting room where they wait for an available bed in the treatment area (again, patients with certain conditions may follow a different route). In addition, we highlight here that PT can be applied in various ways in practice and our modeling approach enables us to capture these differences (see §3.2 for more details).

The general patient flow described above is based mainly on our observations in an ED that we collaborated with (referred to as ED X throughout), but the patient flow and treatment decisions in other EDs in the U.S. are also very similar (Rogg et al. 2013, Soremekun et al. 2012). Each patient who arrives in an ED is assigned an acuity level –usually by a nurse– using the popular five-level Emergency Severity Index (ESI) (McHugh et al. 2012). The ESI uses a scale of 1 to 5, where 1 is the most severe and 5 is the least severe (see Gilboy et al. 2011). In many EDs, including ED X, the most severe cases (ESI Level 1) and the least severe cases (ESI Level 4 and 5) are treated in separate areas in the ED with their own dedicated staff and resources (referred to as "Trauma Bay" and "Fast Track," respectively, in ED X). In such systems, patients with ESI Level

2 and 3 are most severely affected by ED crowding because they because they have to wait for a treatment bed before being treated. Therefore, we mainly consider the triage method decisions regarding patients with ESI Level 2 and 3 in our model and extend our results to more general patient flow structures in the appendix.

The use of PT can change the patient flow and affect the performance in EDs, as discussed widely in the medical literature (see Oredsson et al. 2011 for a review of these studies). PT leads to shorter door-to-initial-provider evaluation times, known as door-to-doctor time, as patients have contact with the provider sooner (door-to-doctor time is one of the crucial metrics recorded by EDs because of its impact on very severe cases such as myocardial infarction). Therefore, fewer patients leave the ED without being seen by a provider when PT is applied (Subash et al. 2004, Holroyd et al. 2007, Han et al. 2010). Also, a triage provider is authorized to order a number of additional tests that a triage nurse cannot order. This potentially leads to more diagnostic tests being ordered during PT and fewer tests needed in treatment rooms. Thus, a patient would spend less time in a treatment bed (Choi et al. 2006). As discussed above, some treatments may be completed by the triage provider. Therefore, fewer patients end up needing treatment beds, and those who are assigned treatment beds spend less time there, which results in higher treatment capacity per bed. Because treatment beds are usually the sources of bottlenecks in EDs (Olshaker and Rathlev 2006), PT can also be utilized to increase overall ED capacity. For example, the application of PT in Scripps Mercy Hospital doubled the ED's capacity without adding additional beds, reduced waiting times from five hours to two hours, and cut left-without-being-seen (LWBS) rates from 8% to 2% (Clark 2010). However, PT is not free. Staffing costs may increase under PT due to potential changes in the ED staffing level. In addition, because the provider may start the treatment of a patient during triage and the hospital may not be fully reimbursed by the healthcare payer for the cost of treatment if the patient abandons the ED, the cost of an abandoning patient may be higher under PT.

Although the benefits of PT are documented in several empirical studies, the extant literature lacks an analytical approach for choosing the triage method during the course of a day in EDs. The general practice is to deploy PT when the patient volume is "high" (see Holroyd et al. 2007 and Han et al. 2010). Our goal in this study is to gain insight into *when* to apply PT in an ED, based on certain economic considerations by comparing the system performances under PT and NT, and also to assess if the current common practice of PT is sensible.

We investigate these issues with the help of a queueing model we develop to capture the effects of each triage method on patient flow. Because the exact analysis of this queueing model does

not provide practical insights, we use a many-server fluid approximation (see Whitt 2004). The analysis of this model in the setting of ED X shows that NT or PT may be preferred depending on the arrival rate at the ED. We show that NT always outperforms PT when the arrival rate is sufficiently low, but PT can outperform NT as the arrival rate increases. However, when the arrival rate becomes sufficiently high, NT may be preferred once again because of, for example, the potentially higher abandonment cost per patient under the PT method or limited PT capacity.

Finally, we test the performance of our proposed policy, which is developed from the aforementioned approximations, by using operational data collected between March 2011 to May 2012 from ED X, which receives 8,000 arrivals per month on average. We first simulate the system for 24 hours and identify the best triage method in one-hour blocks using total enumeration. We then compare this triage method to the one suggested by our methodology under several scenarios. The results indicate that the performance of our solution method is remarkably close to that of the best solution obtained via simulation, specifically, it only leads to a 0.32% decrease in the objective on average compared to the optimal triage method, with a maximum of 0.82% and median of 0.22%. We also find that not utilizing PT and NT effectively could degrade the objective as much as 10%. Also, our solution methodology is computationally much more efficient, as triage decisions in an ED can be made using back-of-the-envelope calculations as opposed to using simulations and total enumeration, which takes approximately four months using a standard PC for each 24-hour period.

The rest of the paper is organized as follows: In §2 we review the related medical and operations management literature. In §3 we describe the setting of our models and the process in the ED when NT and PT are applied and define the objective function of our problem. In §4 we discuss the solution procedure, which is based on steady-state many-server fluid approximations and discuss several extensions of our base model. In §5 we explain the structure of the proposed policy for the implementation of PT in ED X. We conclude the paper in §6 with a summary of our findings, managerial insights, and limitations of our work.

2. Literature Review

There are several comprehensive reviews of the alternative triage system implementations in the medial literature (see Gilboy et al. 2011 among others). Alternative triage methods have also been studied in the operations management literature (see Saghafian et al. 2015 for an excellent review of these studies). Among these alternative applications, we discuss the literature on PT as it is our main focus.

The literature on PT mostly consists of empirical before-after studies. We begin by summarizing the literature that analyzes the effect of PT on patient flow times. Han et al. (2010), Holroyd et al.

(2007), and Traub et al. (2015) show that PT leads to a shorter average ED length of stay. A major reason for this, according to Chan et al. (2005), is the reduction in waiting times. Shorter wait times due to PT are also shown in the simulation-based studies of Holm and Dahl (2009) and Travers and Lee (2006). Choi et al. (2006) report a reduction of 38% in average wait time and 23% in average treatment time. They claim that the wait and processing times of low-acuity patients who were not triaged by a provider during PT intervention were also improved due to the more efficient processing of urgent patients who were triaged by a provider.

Soremekun et al. (2012) and Subash et al. (2004) also show a reduction in time to initial provider evaluation and time to radiology following the application of PT. Burström et al. (2012) compare the reduction in time until first treatment and ED length of stay in an ED under physician-led team triage, in which the treatment was performed by a physician and a junior physician, and two types of NT methods. They show that physician-led team triage outperformed the other two methods.

Han et al. (2010) and Holroyd et al. (2007) show a decline in LWBS rates, and Terris et al. (2004) show that the number of patients waiting to be seen decreases when PT is applied. Rogg et al. (2013) bring up a new and crucial aspect of PT: discharging patients without having to use monitored (or treatment) beds. They show that 18% of patients were discharged without using monitored beds in the first six months of PT intervention. The study by Soremekun et al. (2012) differs from other before-after studies discussed so far in the sense that it combines the operational and financial aspects of PT. Their two-year before-after study provides insight as to what the operational effects of PT might be and estimates a 13-month break-even time from the initial PT investment of $1,200,000 to create four screening rooms and a post-screening internal waiting area. However, none of these studies tries to find out when it is appropriate to use PT.

There is growing interest in operations management literature to examine the reasons and the impact of growing congestion on patient outcomes in EDs. Using operational data from EDs, Song et al. (2015) examine the impact of different patient assignment procedures to physicians on time in ED, Batt and Terwiesch (2016) study the impact of congestion on ED operations and how it affects time in ED by focusing on early task initiation, Kc (2013) studies the impact of multi-tasking on physician productivity, and Allon et al. (2013) examine the factors that drive ambulance diversion. Using analytical models, Xu and Chan (2016) examine how information about future arrivals can be used to reduce congestion, Dobson et al. (2013) propose physician assignment policies to reduce abandonments in the ED, and Saghafian et al. (2012) demonstrate that streaming patients into two groups at the triage step based on their likelihood of being admitted may reduce congestion

in the ED. However, these papers do not examine how alternative triage staffing methods affect congestion.

More relevant to our work in the operations literature are Zayas-Cabán et al. (2014) and Huang et al. (2012), both of which analyze patient flows in EDs and consider triage steps explicitly. Zayas-Cabán et al. (2014) model the ED as a two-step service system where the first step is triage and the second step is treatment, a model we adopt. They assume that two steps are carried out by the same provider and the main decision is how to prioritize the work of the provider based on delays. However, in our model, the providers in two steps are different (based on our observations in practice), hence the assignment problem does not arise in our context. Huang et al. (2012) also study how to allocate physician capacity among patients at different stages of treatment in EDs. In their model, physicians in the treatment area serve two patient groups. The first group comprises patients who are waiting for their initial provider evaluation after triage and are thus on a deadline due to the time-until-first-provider-contact requirement, and the second group comprises patients whose treatment has already been initiated by a provider (in-process, or IP, patients). In our model, if PT is applied, the initial provider evaluation of all patients is conducted by the triage provider, and providers in the treatment area serve only IP patients.

Shumsky and Pinker (2003) also consider a two-step service system in which the steps are carried out by a gatekeeper and a specialist. In their study, the gatekeeper has control over the amount of effort put into service, and the main focus is on how to incentivize the gatekeeper to choose the optimal effort for the firm. However, unlike the ED setting we are interested in, they assume both steps have unlimited capacity and that the amount of effort put into treatment in the first stage does not affect the service times in the second step.

3. Model Description and Motivation

In this section, we present the details of our model and objective function. First, we provide a quick description of different ways PT is implemented in practice in §3.1. We then present our queueing models motivated by these implementations in §3.2, and the objective function in §3.3.

3.1. Provider Triage in Practice

Assigning providers to triage in EDs is a relatively new practice, and so there is no consensus on the best way PT can be implemented. For notational and expositional simplicity, we present our solution methodology based on the implementation in ED X. However, our model is flexible and can incorporate the documented differences in practice, as we highlight throughout the paper. Before we present the specific details of our model, we first discuss the differences in the practical

implementation of PT to pave the way for the subsequent discussion. We list the model extensions that incorporate these differences in §4.3 after presenting our model. We should emphasize that we are not saying the implementation in ED X is preferred to alternative implementations. We simply chose it as a base model to drive the analysis because of our experience with this implementation. In developing our model and solution approach, we focus on the so-called "Rapid Medical Evaluation" (RME) provider triage. This is the implementation that we observed in ED X, and we believe that the model for RME is general enough to be used to model other applications of PT. In RME provider triage, the triage provider is authorized to both treat and discharge patients with minor complaints from the PT area as well as to initiate the treatment of more severe patients and place them in the queue for a treatment bed (ACEP 2006, Chan et al. 2005). There are two other common implementations of PT: "See and Treat" and "PT for Severe Patients."

Under See and Treat (also referred to as "Rapid Triage and Treatment (RTT)" or "Triage, Treat and Release (TTR)" in the literature Rogers et al. 2004, Zayas-Cabán et al. 2014), the triage provider only treats patients with minor complaints and discharges them from the triage area. Under the "PT for Severe Patients" model, the triage provider is dedicated to the initial assessment of the severely ill patients who would require further examination in the treatment area (He 2013). Extensions of our model to these other two applications are discussed in §4.3.

In different applications of PT, which patients are seen by a provider at triage also varies. As described above, patients are triaged into five different severity levels in EDs. In ED X and in most relatively large EDs, patients with ESI Levels 1, 4 and 5 follow a different treatment path from those with ESI Level 2 and 3, and only ESI Level 2 and 3 patients are treated by a provider in triage. This will be our *base* model. In other implementations of PT, a provider at triage can also treat ESI Level 4 and 5 patients (see Imperato et al. 2012). Our models and solution approach can be extended to study these applications of PT, as will be discussed in §4.3.

3.2. Models for ED Triage Methods

In order to compare the two triage methods, we use queueing models with two stages for NT and three stages for PT. When NT is applied, all patients first undergo *the NT step*, which is carried out by the triage nurse. Then, after potentially waiting for an available bed, patients undergo *the treatment step* where they occupy a *treatment bed* until they are discharged. When PT is applied, all patients are first triaged by a nurse at the NT step and additionally by a provider at the PT step, and then they undergo the treatment step.

We divide patients into two groups, *simple* or *complex*, based on the resources they require for treatment completion. Simple patients are assumed to be ambulatory in the sense that their treatment can be completed while they are in the waiting room; hence, the evaluation by the triage

provider is sufficient to complete their treatment, and they do not need to be assigned a treatment bed. Complex patients, on the other hand, need to be assigned a treatment bed and require further examination in the treatment area, even when they are triaged by a provider. Typically, ESI Level 2 patients are more severe, and their treatment requires a thorough examination beyond the initial provider evaluation. Hence, we model all ESI Level 2 patients as complex patients. ESI Level 3 patients are less urgent than ESI Level 2 patients and can be either simple or complex patients, for example, in ED X 18% of ESI Level 3 patients are simple patients. (The classification of patients as simple or complex based on their ESI levels may not always be followed strictly as we observed in ED X.) We next discuss the details of our queueing models for NT and PT.

Nurse Triage: The patient flow under the NT method is illustrated in Figure 1. We assume that all patients are triaged by a nurse at the NT step without delay; and hence, there is ample capacity at this step. This assumption is based on two real-life observations. First, door-to-triage times in EDs are typically very short mainly because every patient seeking treatment in an ED must be triaged shortly after their arrival in order to identify life-threatening conditions such as myocardial infarction, seizures, or severe asthma immediately. For example, Subash et al. (2004) report wait times of two to seven minutes for triage. In ED X, wait times were between 5.03 minutes and 9.63 minutes, including time it takes to register the patient. Also, abandonment prior to triage is negligible. (Only 0.72% of patients abandoned the ED before triage, compared to an overall abandonment rate of 5.28% for ESI Level 2 and 3 patients.) Second, triage takes a significantly shorter time than treatment: a median of 1.80 minutes compared to 5.70 hours in ED X, respectively.

After triage, patients proceed to the treatment step. In our model, we use three separate treatment bed queues according to patient severity. ESI Level 2 patients have the highest (non-preemptive) priority; ESI Level 3 complex patients have higher (non-preemptive) priority than ESI Level 3 simple patients. (We assume that complex ESI Level 3 patients have priority over simple patients based on our observations in ED X. Our analysis can easily be extended to the case where all ESI Level 3 patients are served on a first-come-first-served basis, see §4.3.) Within each queue patients are served on a first-come-first-served (FCFS) basis. Also, patients are assumed to have limited patience and will abandon the queue if their waiting time for a treatment bed exceeds their patience time.

Provider Triage: The patient flow under (RME) PT is illustrated in Figure 2. All patients are first triaged by a nurse at the NT step. We assume in our base model that all patients are directed by the triage nurse to the PT step and consider other alternatives in §4.3. Patients are seen at

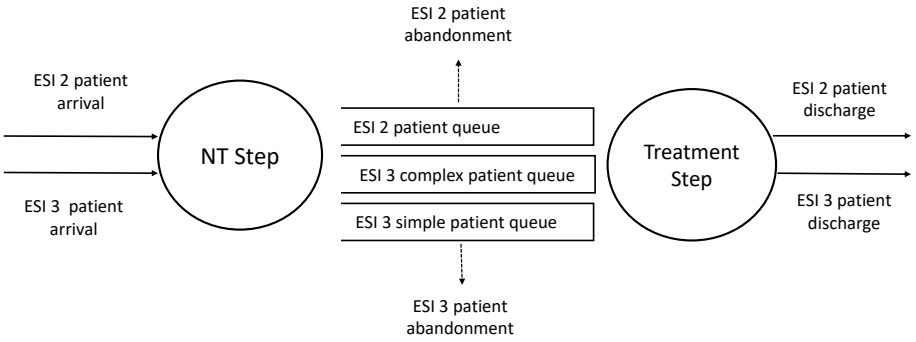


Figure 1 Patient flow when the NT method is applied.

the PT step on a FCFS basis (see Remark 2 for extensions to other priority rules). Also, due to limited capacity in the PT step, if patients who are directed to the PT step wait longer than their patience time, they will abandon the ED. Under PT, simple patients are discharged after the PT step, and all other patients are placed in the queue for a treatment bed, where ESI Level 2 patients are given (non-preemptive) priority over ESI Level 3 complex patients.

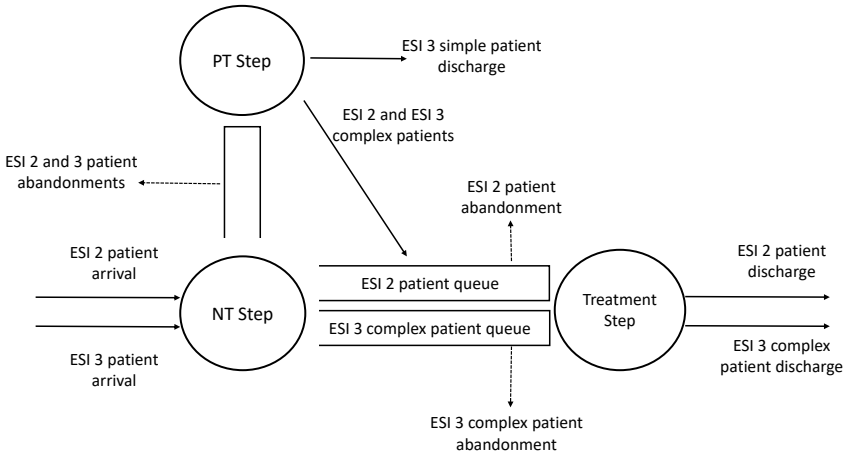


Figure 2 Patient flow when the PT method is applied.

3.3. Model Parameters and Objective

We next define the model parameters and the objective function. We use indices N and P to denote the triage methods NT and PT, respectively. We refer to ESI Level 2 patients as type 1 patients, ESI Level 3 complex patients as type 2 patients, and ESI Level 3 simple patients as type 3 patients for notational simplicity.

The arrival rate at the ED per unit time at time t is denoted by $\lambda(t)$. The proportion of type i patients at time t is denoted by $\gamma_i(t)$, $i = 1, 2, 3$. Thus, the arrival rate of type i patients at

time t, denoted by $\lambda_i(t)$, is $\gamma_i(t)\lambda(t)$. Let $M_j(t)$ denote the number of (staffed) treatment beds allocated to the patient group being considered under the triage method j, $j \in \{N, P\}$. (The number of treatment beds $M_j(t)$ might depend on the triage method due to potential changes in bed allocations and staffing levels under different triage methods.) Finally, the rates at which type i patients are triaged in the PT step and treated in the treatment step under method j are denoted by δ_i and μ_{ij}, respectively, $i \in \{1, 2, 3\}$, $j \in \{N, P\}$.

We use r_{ij}, referred to as net revenue, to denote the revenue that the ED earns for treating a type i patient who is triaged under method j net of the variable treatment cost. The variable cost of treatment can depend on the triage method due to potential changes in the process. However, fixed costs that are the same regardless of the triage method do not have to be accounted for since they do not affect triage method decisions, and triage method-dependent fixed costs associated with additional resources –for example, additional supporting staff– can be accounted for in the staffing cost defined below. For a type i patient, we denote the cost of abandoning the queue for the PT step and the queue for a treatment bed when triage method j is used by y_i and w_{ij}, respectively. We allow the cost of abandoning treatment bed queues to depend on the triage method adopted for the patient because the hospital could potentially incur a cost for the interventions applied at triage in case the patient abandons the treatment queue. We use c_N and c_P to denote the staffing cost for the triage and treatment areas allocated to the patient group being considered under triage methods NT and PT, respectively.

Remark 1. Patient abandonment is costly for the ED based on our experience. However, if the ED is reimbursed for abandoned patients, the total reimbursement for these patients may be included in the total revenue function by adjusting the costs y_i and w_{ij}.

Objective function: Our goal in this setting is to determine when to apply PT in order to maximize the ED's objective for a fixed time interval $[0, T]$. For example, a typical time interval of interest in an ED is from 8 a.m. to 12 a.m., during which the majority of patients –around 85% in ED X– arrive at the ED. Let π denote a triage method policy, and for notational simplicity, we also define π as a stochastic process so that

$$\pi(t) = \begin{cases} P, \text{ if PT is applied at time } t, \\ N, \text{ otherwise.} \end{cases}$$

To avoid subtle technical difficulties, we only consider policies that are Markovian; that is, the triage decisions are made based on the current state of the system. We refer to such triage policies as admissible policies.

The objective function consists of three components. The first part is the revenue earned from patients treated in the ED. We define $S^{\pi}_{ij}(t)$ and $D^{\pi}_{ij}(t)$ to denote the number of discharges before the treatment step (i.e., discharges from the PT step) and after the treatment step, respectively, until time t for type i patients who have been triaged under method j. The second part is the abandonment cost. We denote the number of patients who abandon the ED until time t before joining the treatment bed queue and while queueing for a treatment bed by $B^{\pi}_{ij}(t)$ and $E^{\pi}_{ij}(t)$, respectively, for type i patients who have been triaged under method j. The last component is the staffing cost based on the length of time for which each triage method is used and is given by

$$K^{\pi}(t) = c_N \int_0^t \mathbb{1}\{\pi(s) = N\}\,\mathrm{d}s + c_P \int_0^t \mathbb{1}\{\pi(s) = P\}\,\mathrm{d}s, \tag{1}$$

where $\mathbb{1}$ is the indicator function. Our objective is to find a triage policy π that maximizes

$$\Phi^{\pi}(T) = \sum_{i=1}^{3} \sum_{j\in\{N,P\}} (r_{ij}(\mathbb{E}[D^{\pi}_{ij}(T)] + \mathbb{E}[S^{\pi}_{ij}(T)]) - y_i \mathbb{E}[B^{\pi}_{ij}(T)] - w_{ij}\mathbb{E}[E^{\pi}_{ij}(T)]) - K^{\pi}(T). \tag{2}$$

We also consider an alternative objective where the goal is to minimize the number of abandonments; see §4.3 for more details.

4. Solution Methodology

In this section, we describe our solution methodology which is based on fluid approximations. In order to evaluate (2) analytically for a fixed policy, we need to determine $\mathbb{E}[D^{\pi}_{ij}(T)]$, $\mathbb{E}[E^{\pi}_{ij}(T)]$, $\mathbb{E}[B^{\pi}_{ij}(T)]$ and $\mathbb{E}[S^{\pi}_{ij}(T)]$ under different policies. However, closed-form solutions cannot be obtained even under trivial policies. Also, when the policy is time-dependent, it is necessary to keep track of the triage method applied for each patient (i.e., the triage method in use at each patient's time of arrival) in addition to the patient type. This makes obtaining exact solutions even more unlikely. Therefore, to obtain a triage policy that can be easily determined and that provides additional insight, we use fluid approximations. We explain the details of these approximations in §4.1. In §4.2, we present our solution method. Finally, in §4.3 we discuss the extensions of our model.

4.1. Fluid Approximations

Our approximations are based on the pointwise stationary approximations in Green and Kolesar (1991) and Green et al. (1991). Under these approximations, the system is assumed to reach steady-state instantaneously at each point in time. We approximate the steady-state of the system for a fixed arrival rate using fluid approximations.

Consider an ED model with a fixed arrival rate λ, fraction of arrivals γ_i and arrival rate λ_i for patient type i, where $\lambda_i = \gamma_i\lambda$, $i = 1, 2, 3$. We denote by M_j the (staffed) treatment beds allocated

to patients undergoing treatment. We use $s_{ij}(\lambda)$ and $d_{ij}(\lambda, M_j)$ to denote the rate of discharge of type i patients from triage and treatment, respectively, under triage method j. Similarly, we denote the rates of abandonment from the triage queue and the treatment bed queue by $b_{ij}(\lambda)$ and $a_{ij}(\lambda, M_j)$, respectively. Finally, we denote the rate at which type i patients arrive at the treatment step queue by $\kappa_{ij}(\lambda)$ under triage method j.

In the rest of this section, we discuss how to approximate the terms defined above using fluid limits under NT and PT methods.

4.1.1. Nurse Triage: By our assumption of ample capacity at the NT step, patients are triaged by the nurse without delay and are placed in treatment bed queues. Hence, the arrival rate of type i patients at the treatment bed queue is

$$\kappa_{iN}(\lambda) = \lambda_i, \ i = 1, 2, 3. \tag{3}$$

Because type 1 patients have the highest priority in treatment bed assignment, the number of treatment beds they occupy is

$$M_{1N}(\lambda, M_N) = \min\left\{M_N, \frac{\kappa_{1N}(\lambda)}{\mu_{1N}}\right\}. \tag{4}$$

Because type 2 patients have priority over type 3 patients, the remaining $(M_N - M_{1N}(\lambda, M_N))$ treatment beds are available to type 2 patients. Thus, the number of beds occupied by type 2 patients is

$$M_{2N}(\lambda, M_N) = \min\left\{M_N - M_{1N}(\lambda, M_N), \frac{\kappa_{2N}(\lambda)}{\mu_{2N}}\right\}. \tag{5}$$

Finally, the number of treatment beds occupied by type 3 patients is

$$M_{3N}(\lambda, M_N) = \min\left\{M_N - M_{1N}(\lambda, M_N) - M_{2N}(\lambda, M_N), \frac{\kappa_{3N}(\lambda)}{\mu_{3N}}\right\}. \tag{6}$$

Hence, the rate of discharge from treatment and the rate of abandoning treatment bed queues are respectively given by

$$d_{iN}(\lambda, M_N) = \mu_{iN} M_{iN}(\lambda, M_N), \quad a_{iN}(\lambda, M_N) = \kappa_{iN}(\lambda) - d_{iN}(\lambda, M_N), \ i = 1, 2, 3. \tag{7}$$

4.1.2. Provider Triage: Recall that under the PT method, all patients first go through the NT step and then are placed in the PT step queue. In general, patients are evaluated at the PT step on a FCFS basis; hence, the average service rate at the PT step is $\left(\sum_{i=1}^{3} \frac{\gamma_i}{\delta_i}\right)^{-1}$. Then, the rate at which type i patients abandon the PT step queue is

$$b_{iP}(\lambda) = \gamma_i \left(\lambda - \left(\sum_{i=1}^{3} \frac{\gamma_i}{\delta_i}\right)^{-1}\right)^{+}, \ i = 1, 2, 3, \tag{8}$$

where $x^+ = \max\{x, 0\}$. Once their interventions at the PT step are completed, type 1 and 2 patients are placed in line for treatment bed assignment, whereas type 3 patients are discharged from the ED. Hence, the rates of discharge before being placed in treatment bed queue are

$$s_{iP}(\lambda) = 0 \text{ for } i = 1, 2, \quad s_{3P}(\lambda) = \lambda_3 - b_{3P}(\lambda). \tag{9}$$

The rate at which type i patients join the treatment bed queue is therefore given by

$$\kappa_{iP}(\lambda) = \lambda_i - b_{iP}(\lambda) - s_{iP}(\lambda), \ i = 1, 2, 3. \tag{10}$$

Because type 1 patients are given priority over type 2 patients in treatment bed assignment and by $\kappa_{3P}(\lambda) = 0$, the number of treatment beds occupied by each patient type is given by

$$\begin{aligned} &M_{1P}(\lambda, M_P) = \min\left\{M_P, \frac{\kappa_{1P}(\lambda)}{\mu_{1P}}\right\}, \quad M_{2P}(\lambda, M_P) = \min\left\{M_P - M_{1P}(\lambda, M_P), \frac{\kappa_{2P}(\lambda)}{\mu_{2P}}\right\}, \\ &M_{3P}(\lambda, M_P) = 0. \end{aligned} \tag{11}$$

Thus, the rate at which type i patients are discharged from the treatment area is

$$d_{iP}(\lambda, M_P) = \mu_{iP} M_{iP}(\lambda, M_P), \tag{12}$$

and the rate at which type i patients abandon the treatment bed queue is

$$a_{iP}(\lambda, M_P) = \kappa_{iP}(\lambda) - \mu_{iP} M_{iP}(\lambda, M_P), \ i = 1, 2, 3. \tag{13}$$

Remark 2. The PT method can alternatively be implemented by prioritizing patients based on their severity at the PT step. In this case, type 1 and 3 patients would have the highest and lowest priority, respectively, similar to prioritization in the treatment bed queue. Under this severity-based prioritization, the rate at which type i patients abandon the PT step queue is

$$b_{1P}(\lambda) = (\lambda_1 - \delta_1)^+, b_{2P}(\lambda) = \left(\lambda_2 - \left(1 - \frac{\lambda_1}{\delta_1}\right)^+ \delta_2\right)^+, b_{3P}(\lambda) = \left(\lambda_3 - \left(1 - \frac{\lambda_1}{\delta_1} - \frac{\lambda_2}{\delta_2}\right)^+ \delta_3\right)^+. \tag{14}$$

By using b_{iP} as defined in (14), the outcomes under PT method can be approximated by (9)–(13).

Remark 3. PT can potentially lead to additional costs for some of the patients as it introduces a handover from the triage provider to the provider in the treatment area (see Ye et al. 2007 and Cheung et al. 2010 for more on handovers in EDs). Because we model net revenue per patient as being dependent on the triage method applied on the patient, our approach can deal with handover costs by simply subtracting the handover costs due to PT from the net revenue per patient and by taking into account the effect of handovers on the treatment time.

4.2. Proposed Solution

In this section, we present our approximations for the objective function $\Phi^{\pi}(T)$. Let $\Theta^{j}(\lambda, M_j)$ denote the rate of change of the objective function per unit time in steady-state when the arrival rate is λ, the triage method is j, and the number of (staffed) treatment beds is M_j, $j \in \{N, P\}$. Based on the prescribed fluid approximations, we arrive at the following approximations,

$$\begin{aligned}\Theta^{N}(\lambda, M_N) &= \sum_{i=1}^{3} r_{iN} d_{iN}(\lambda, M_N) - \sum_{i=1}^{3} w_{iN} a_{iN}(\lambda, M_N) - c_N, \\ \Theta^{P}(\lambda, M_P) &= \sum_{i=1}^{3} r_{iP}(d_{iP}(\lambda, M_P) + s_{iP}(\lambda)) - \sum_{i=1}^{3} (y_i b_{iP}(\lambda) + w_{iP} a_{iP}(\lambda, M_P)) - c_P.\end{aligned} \tag{15}$$

Using (15), we arrive at the following approximation $\hat{\Phi}^{\pi}(T)$ for the objective function Φ^{π} under policy π

$$\hat{\Phi}^{\pi}(T) = \left[\int_0^T \Theta^{\pi(t)}(\lambda(t), M_{\pi(t)}(t)) dt \right]. \tag{16}$$

One of the important features of our approximations is that at each time point the approximation is independent of the state of the system prior to that point, because they are based on steady-state quantities. Therefore, the triage method decision can be made in isolation at each time point. We define

$$j^*(\lambda, M_N, M_P) = \begin{cases} N, \text{ if } \Theta^{N}(\lambda, M_N) \geq \Theta^{P}(\lambda, M_P), \\ P, \text{ otherwise.} \end{cases} \tag{17}$$

Our *proposed solution* is to use triage method $j^*(\lambda(t), M_N(t), M_P(t))$ at time t.

4.3. Extensions

So far our main focus has been the version of the PT implemented at ED X with a few extensions discussed above for some alternative implementations. Before we proceed to demonstrate the application of our results in ED X in the next section, we emphasize the fact that our modeling approach is flexible. We consider several additional aspects of PT as extensions of our base model in Appendix **??** and summarize them here. (i) In certain applications of PT, ESI Level 4 and 5 patients are also triaged by the provider. We present the extension of our model to this case in Appendix **??**. (ii) ED managers may choose to allocate the limited PT step capacity to only a fraction of patients. This alternative implementation of PT represents a mix of the NT and PT models that we consider in our base model and is examined in Appendix **??**. We also explain how "See and Treat" and "PT for Severe Patients" implementations can be analyzed using this extension. Additionally, we provide a method to determine the optimal fraction of patients who should

be directed to the PT step. (iii) The misclassification of patients in triage has been documented in the literature (see Saghafian et al. 2014). We extend our model to account for misclassifications in Appendix **??**. (iv) Also, because the PT step has limited capacity, it is not clear whether the PT method would decrease the number of abandonments from the ED. In Appendix **??**, we provide an alternate objective function of minimizing the number of abandonments. (v) In Appendix **??**, we discuss how to use our solution methodology in EDs that prioritize patients only according to their acuity level (i.e., ESI level) in the treatment bed queue, that is, when complex and simple patients in the same ESI level are treated in a FCFS manner.

5. The Implementation of the Proposed Policy in ED X

In this section, we apply the solution approach in §4 in the setting of ED X. We have three primary goals: (i) to demonstrate the application of the proposed method, (ii) to gain insight on the implementation of PT under simplifying assumptions, and (iii) to prove that the proposed method is asymptotically optimal. In §5.1 we present the details of how PT is implemented in ED X. In §5.2 we examine the proposed solution for PT practice in ED X and provide insights on implementation in §5.3. In §5.4 we show that our solution method for ED X is asymptotically optimal in large systems in a certain asymptotic regime, and in §5.5 we present the results of numerical experiments to assess the effectiveness of the proposed solutions.

5.1. PT Practice in ED X

As described in §3.1, PT is implemented in different ways in practice. In this section, we describe the implementation in ED X and how this implementation can be analyzed using our method. In ED X, an additional provider is added to the triage step without changing the staffing level or the number of treatment beds in the treatment area. (We analyze the case where instead a provider is assigned from the treatment area to the triage area in Appendix **??**.) Hence, the staffing cost is higher under the PT method, that is, $c_P > c_N$.

Based on the implementation details in ED X, we make two simplifying assumptions in our model. First, we only consider one type of simple patients and assume that the revenue per discharged patient is the same under both triage methods. The rationale behind the first assumption is based on the fact that almost all of the ESI Level 2 patients bypass the PT step and queue for a treatment bed immediately after triage. (Recall that ESI Levels 1, 4, and 5 follow a different treatment path.) Our second assumption is based on the observation that the triage method does not affect the reimbursement levels at all and does not have a significant impact on treatment costs. These two assumptions simplify the analysis and allow us to gain insight on the triage decisions. Even

when these assumptions do not hold exactly, the insight from our analysis should still be valid if deviations are small because our approximations for the revenue functions under both triage methods are linear (discussed in §5.2).

We also make the following additional assumptions. We assume that the number of (staffed) treatment beds is the same under both triage methods, that is, $M \equiv M_N = M_P$. We also observed that patients rarely wait before the PT step in ED X. This is because patients are placed in the waiting room after the brief evaluation by the triage provider and do not occupy resources (e.g., the triage room) in the PT step until the triage provider receives their test results. Hence, we mainly focus on the case when the PT step has unlimited capacity and later also analyze the case where it is limited. We assume that the abandonment cost for complex patients is higher than that for the simple patients, that is, $w_{2j} \geq w_{3j}$ for $j \in \{N, P\}$. Also, fewer tests are ordered at the triage stage when NT is used because triage nurses are only authorized to order a limited set of basic tests. Therefore, we assume that $w_{iN} \leq w_{iP}$ for $i \in \{2, 3\}$. We also assume that $\mu_{2P} > \mu_{2N}$, that is, the complex patients are treated more quickly under PT than NT (see Appendix **??** for the analysis of the case with $\mu_{2P} \leq \mu_{2N}$). Because treatment procedures for complex patients are more intensive, we assume that $\mu_{3N} \geq \mu_{2N}$ and (dropping the triage method subscript in the net revenue term for the rest of this section) $r_2 \geq r_3$. We denote the fractions of type 2 and type 3 patients by $1-\gamma$ and γ, respectively.

5.2. Proposed Policy for ED X

We next examine the proposed triage method policy for practice in ED X. We show below that there are three possible structures of the optimal triage method, j^*, based on the values of the parameters as λ grows larger for constants $\Lambda_1 < \Lambda_2$ that depend on the system parameters:

1. $j^*(\lambda, M) = N$ for all $\lambda \geq 0$. Figure 3(a) provides an illustration. (Piecewise linearity follows from the definitions of Θ^N and Θ^P);
2. $j^*(\lambda, M) = N$ for all $\lambda \leq \Lambda_1$ and $j^*(\lambda, M) = P$ for all $\lambda > \Lambda_1$. Figure 3(b) provides an illustration;
3. $j^*(\lambda, M) = N$ for all $\lambda \leq \Lambda_1$ and $\lambda \geq \Lambda_2$, $j^*(\lambda, M) = P$ for $\Lambda_1 < \lambda < \Lambda_2$. Figure 3(c) provides an illustration.

Which solution prevails can be determined based on certain conditions listed below that the parameters satisfy. Because the objective functions Θ^N and Θ^P are piecewise linear and $\Theta^N(0, M) > \Theta^P(0, M)$ (since $c_N < c_P$), comparing the values of these functions at their break points and comparing their slopes above the largest break point are sufficient to identify their intersection.

$$\textit{Condition 1: } \gamma(r_3 + w_{3N})\frac{M\mu_{2N}}{(1-\gamma)} \geq c_P - c_N. \tag{18}$$

$$\textit{Condition 2: } (r_2+w_{2N})M(\mu_{2P}-\mu_{2N})+(r_3+w_{3N})\frac{\gamma M\mu_{2P}}{(1-\gamma)}\geq c_P-c_N. \tag{19}$$

$$\textit{Condition 3: } \gamma(r_3+w_{3N})\geq(1-\gamma)(w_{2P}-w_{2N}). \tag{20}$$

We next present the main result and discuss the main insights in §5.3 below.

PROPOSITION 1. *The optimal solution j^* is given by the following.*

1. *$j^*(\lambda,M)=N$ for all $\lambda\geq 0$, if Conditions 1 and 2 do not hold, and Condition 3 either does not hold or holds as an equality.*
2. *$j^*(\lambda,M)=N$ for all $\lambda\leq\Lambda_1$ and $j^*(\lambda,M)=P$ for all $\lambda>\Lambda_1$, if Condition 3 holds and at least one of the conditions in part (1) does not hold, where*

(a)

$$\Lambda_1=\frac{\mu_{2N}(M\mu_{3N}(r_3+w_{3N})+c_P-c_N)}{(r_3+w_{3N})(\gamma\mu_{2N}+(1-\gamma)\mu_{3N})}, \tag{21}$$

if Condition 1 holds,

(b)

$$\Lambda_1=\frac{(r_2+w_{2N})M\mu_{2N}+c_P-c_N}{\gamma(r_3+w_{3N})+(1-\gamma)(r_2+w_{2N})}, \tag{22}$$

if Condition 1 does not hold but Condition 2 holds,

(c)

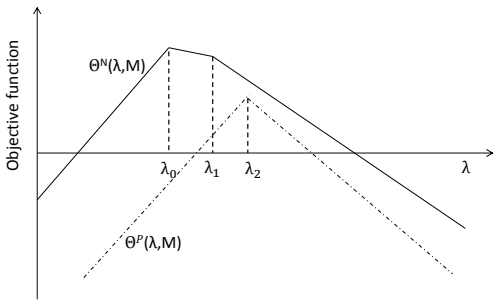
$$\Lambda_1=\frac{(r_2+w_{2N})M\mu_{2N}-(r_2+w_{2P})M\mu_{2P}+c_P-c_N}{\gamma(r_3+w_{3N})-(1-\gamma)(w_{2P}-w_{2N})}, \tag{23}$$

if both Conditions 1 and 2 fail to hold and Condition 3 holds as a strict inequality.

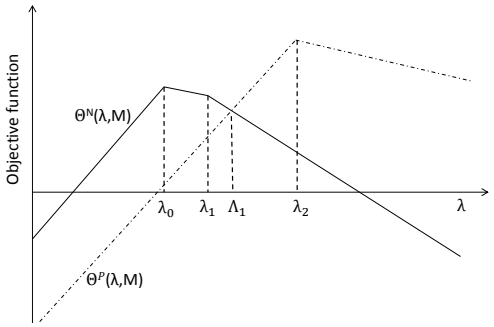
3. *$j^*(\lambda,M)=N$ for all $\lambda\leq\Lambda_1$ and $\lambda\geq\Lambda_2$, $j^*(\lambda,M)=P$ for $\Lambda_1<\lambda<\Lambda_2$ if Condition 3 does not hold but at least one of Conditions 1 and 2 holds, where*
   (a) *Λ_1 is given by (21), and Λ_2 is given by the right-hand side of (23) if Condition 1 holds,*
   (b) *Λ_1 is given by (22), and Λ_2 is given by the right-hand side of (23) if Condition 1 does not hold but Condition 2 does.*

The triage method policy suggested by Proposition 1 is referred to as *the threshold policy*. We provide a sketch of the proof below and its details in Appendix **??**.

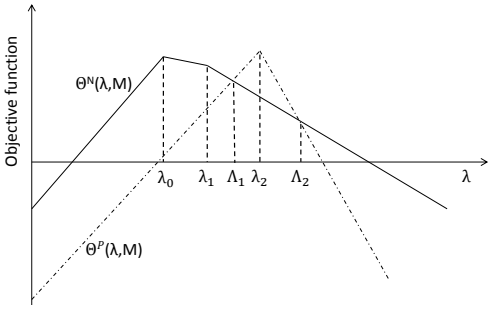
Sketch of the proof of Proposition 1: Figure 3 illustrates how the objective functions Θ^N and Θ^P depend on the arrival rate and depicts the scenarios explained in Proposition 1. First, when the arrival rate is low, NT is preferable to PT under all scenarios as explained above. As the arrival rate increases, the objective functions for both systems increase until the systems become



(a) No Intersection.



(b) One Intersection.



(c) Two Intersections.

Figure 3 $\Theta^N(\lambda, M)$ **and** $\Theta^P(\lambda, M)$ **in ED X vs.** λ **with zero, one or two intersections.**

overloaded. Because the system under NT has a lower capacity, it reaches full capacity (at point λ_0 in Figures 3(a)–(c)) before the system under PT does (at λ_2 in Figures 3(a)–(c)).

Once each system becomes overloaded, some of the patients will abandon the system. For the system under NT, initially simple patients will abandon after the arrival rate exceeds λ_0 because complex patients have priority. When the arrival rate exceeds λ_1, some of the complex patients will abandon as well because there is insufficient capacity to serve all the complex patients. Hence, the slope of Θ^N changes at λ_0 and then again at λ_1. On the other hand, the slope of Θ^P changes only once at λ_2 because all the simple patients are treated at the triage step.

Condition 1 implies that Θ^P is greater than or equal to Θ^N at arrival rate λ_1. At λ_1, all patients can be treated if PT is used, whereas only complex patients are treated under NT. The left-hand side of Condition 1 is the additional revenue and the right-hand side is the additional cost under PT when the arrival rate is equal to λ. Similarly, Condition 2 implies that Θ^P is greater than or equal to Θ^N at λ_2 when all patients are treated if PT is used, but all of the simple and some of the complex patients abandon under NT. Again, the left-hand side of Condition 2 is the additional revenue under PT at this point. Finally, Condition 3 implies that the slope of Θ^P is greater than or equal to that of Θ^N beyond λ_2.

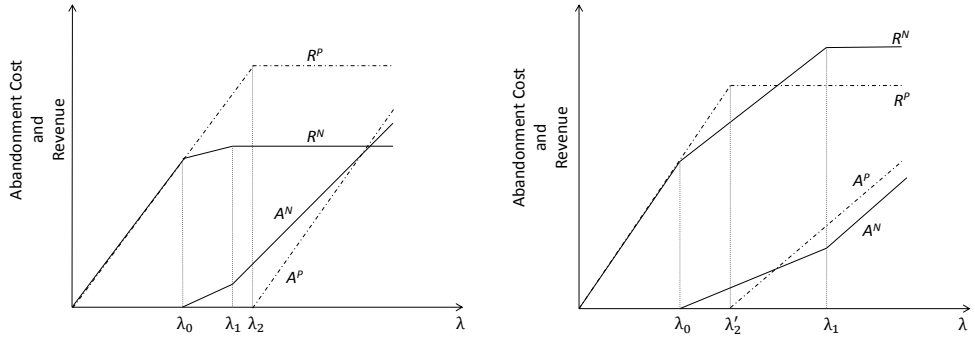
5.3. Insights

Proposition 1 reveals an interesting phenomenon; even when PT becomes more beneficial than NT as arrival rate increases, NT may become more beneficial again if the arrival rate is sufficiently high (see Figure 3(c)). In this section, we first determine the factors that drive this result for the base ED X case and then demonstrate that this result can also hold for alternative implementations of PT discussed earlier.

In Figure 4(a), we present the total revenues R^N and R^P as well as the abandonment costs A^N and A^P under the NT and PT methods, respectively, as a function of the total arrival rate (following the notation introduced in Figure 3). Figure 4(b) is also similar, but it is for the case with limited PT capacity discussed below. The revenue and abandonment costs are computed by

$$R^j(\lambda, M) = \sum_{i=2}^{3} r_i d_{ij}(\lambda, M), \; A^j(\lambda, M) = \sum_{i=2}^{3} w_{ij} a_{ij}(\lambda, M) \text{ for } j = N, P.$$

The plots for the objective functions in Figure 3 are clearly based on combining R^j and A^j in Figure 4 with the staffing costs.



(a) Unlimited PT capacity (b) Limited PT capacity

Figure 4 Abandonment cost and revenue functions under NT and PT.

Figure 4 demonstrates clearly why PT is not always more beneficial than NT when the arrival rate is high. In fact, when the arrival rate is sufficiently high, the revenue under PT is higher than that under NT. However, its abandonment costs are higher as well, which, under certain conditions, offsets its advantage in revenue.

We observe the same phenomenon in the following cases as well: (i) when PT has limited capacity and all patients are directed to PT step (see Appendix **??** for details), (ii) when PT has limited capacity that can be used by only a fraction of patients (see Appendix **??** for details), (iii) when a provider is moved to triage from the treatment step (see Appendix **??** for details), and (iv) for the other alternative implementations of PT discussed in §4.3 (see Appendix **??**). We next explain the

case where PT has limited capacity. Consider the cost and the revenue functions in Figure 4(b) for this case (see Appendix **??** for a detailed analysis). If the total capacity of the PT step is λ_2' and the arrival rate exceeds this threshold, the revenue for PT remains constant beyond this point. In addition, if $\lambda_0 < \lambda_2' < \lambda_1$ and $\Theta^P(\lambda_1) \leq \Theta^N(\lambda_1)$, PT is preferred if the arrival rate is high, but NT is preferred when it is sufficiently higher. In summary, an important outcome of our analysis is that PT should not always be favored over NT when the arrival rate at an ED is high. There are other factors that need to be considered, such as abandonment costs and PT capacity depending on the implementation, and our solution method provides a relatively simple way to incorporate these factors in triage decisions.

5.4. Asymptotic Optimality of the Proposed Solution

In this section, we prove that the methodology in Proposition 1 is asymptotically optimal in large systems. Although we only focus on the implementation in ED X for simplicity, a similar result can be proved for other systems under additional assumptions. We consider an asymptotic regime that is used to study systems with time-varying arrivals (see Bassamboo et al. 2006, Besbes and Maglaras 2009, Stolyar and Tezcan 2011, Pinker and Tezcan 2013 and the references therein). Our goal in this section is twofold: (i) to show the basis of our approximations and prove that they are obtained using an asymptotic regime, and (ii) to obtain insight into when our approximations yield accurate results. In addition, we note that the current analysis is different from the papers listed above due to the fact that there is a discontinuity in the system when the ED switches triage methods. This makes the analysis more complicated. We verify here that the pointwise approximations we use are still valid despite this discontinuity.

We consider a sequence of EDs indexed by n. Let $\{k^n\}$ denote a sequence of real numbers such that $k^n \to \infty$ and $k^n/n \to 0$ as $n \to \infty$ and $T^n = k^n T$. Assume that the arrival rate Λ^n in the n^{th} ED satisfies

$$\Lambda^n(t) = n\Lambda\left(\frac{t}{k^n}\right), \tag{24}$$

for a nonnegative continuous function Λ satisfying

$$\sup_{t\in[0,T]} \|\Lambda(t)\| < c_\Lambda, \tag{25}$$

for some constant $c_\Lambda < \infty$. Therefore, along this sequence, the arrival rate increases but the relative rate of change decreases with n. We also assume that capacity during the treatment stage is scaled similarly such that

$$M^n(t) = nM\left(\frac{t}{k^n}\right), \tag{26}$$

for a nonnegative continuous function M satisfying

$$\sup_{t\in[0,T]} \|M(t)\| < M_c, \tag{27}$$

for some constant $M_c < \infty$. Following (26), we assume that the staffing costs under NT and PT in nth ED have the following form:

$$c_j^n = nc_j,\ j = N, P. \tag{28}$$

We further assume that the number of times the suggested triage method can change in the time interval $[0, T]$ is finite.

ASSUMPTION 1. *The function* $\pi'(t) = \mathbb{1}\{j^*(\Lambda(t), M(t)) = N\}$ *has finitely many discontinuities in* $[0, T]$.

First, we consider the solution for (2) in the nth ED. Let Π denote the set of admissible policies. We set

$$\bar{\Phi}^\pi(T^n) = \frac{\Phi^\pi(T^n)}{k^n n} \text{ and } \bar{\Phi}^*(T^n) = \sup_{\pi\in\Pi} \bar{\Phi}^\pi(T^n).$$

Using (16), we also define

$$\Phi^* = \int_0^T \Theta^{\pi^*(t)}(\lambda(t), M(t))dt, \tag{29}$$

where

$$\pi^*(t) = j^*\left(\Lambda(t), M(t)\right), \tag{30}$$

and j^* is defined in (17). Also, our solution procedure calls for the use of triage policy $\pi^{*,n}(t)$ at time t, which is defined as follows:

$$\pi^{*,n}(t) = j^*\left(\Lambda^n(t), M^n(t)\right). \tag{31}$$

Note that by (24), (26) and (31), the arrival rates at which the suggested triage method changes under the threshold policy in the nth system are $n\Lambda_1$ and $n\Lambda_2$, where Λ_1 and Λ_2 are defined in Proposition 1.

Let Q^n denote the total queue length in the nth system. We assume that

$$\lim_{n\to\infty} \frac{Q^n(0)}{n} = \bar{Q}(0) < \infty \text{ a.s.} \tag{32}$$

The following result proves the asymptotic optimality of our proposed triage policy.

THEOREM 1. *(i) Consider a sequence of systems indexed by n that satisfy* (24)–(28) *and* (32). *Then,*

$$\liminf_{n\to\infty} \bar{\Phi}^{\pi^n}(T^n) \geq \Phi^* \tag{33}$$

under any sequence π^n of admissible policies.
(ii) In addition, if Assumption 1 holds, then

$$\lim_{n\to\infty} \bar{\Phi}^{\pi^{*,n}}(T^n) = \Phi^*. \tag{34}$$

Theorem 1 implies that if the triage method is selected according to (15) and (17), then the properly scaled objective function of the triage policy is optimized asymptotically as the system becomes larger. In addition to providing the optimal triage method, (15) and (17) also provide asymptotically correct estimates of the objective function under each threshold policy by (34).

REMARK 4. The motivation for using the scaling (24) comes from empirical studies in the literature (Green et al. 2006, Saghafian et al. 2012, 2014, Yom-Tov and Mandelbaum 2014, Armony et al. 2015, Shi et al. 2015) as well as our own observation that although the arrival rate in EDs is time-dependent, its value does not change vastly from early morning till early evening. Queueing models with stationary arrival rates have already been shown to yield accurate results for EDs under time-varying arrival rates (see Huang et al. 2012). We also show in our numerical experiments in the next section that our solution methodology provides very accurate results for these systems, with arrival rates estimated from actual arrivals to an ED.

5.5. Case Study

In this section, we present a numerical experiment to test the accuracy of the proposed method in the setting of ED X. We only focus on the implementation of PT in ED X as discussed in §5.1, that is, an additional provider is added to the triage step and the PT step has ample capacity.

Parameter Estimation: When estimating parameter values, we use the results from the literature whenever available, mostly for cost and revenue parameters. However, most of the operational parameters that we use (μ_{ij}, λ, γ, M and θ) are not reported in the literature; therefore, we use the data obtained from ED X.

To estimate the revenue per patient, we use the values in Soremekun et al. (2012) and Henneman et al. (2009). We set the abandonment cost per patient equal to the cost of tests ordered at the triage step and hence ignore "goodwill costs." By using the the direct cost per patient visit as reported in Henneman et al. (2009) and calculating the number of tests ordered at the triage and

treatment steps from ED X data, we obtain estimates of abandonment cost per patient. Finally, we estimate the staffing costs under NT and PT using the national averages of mean hourly wages provided by the Bureau of Labor Statistics (BLS 2012).

For the estimation of operational parameters, we use the data from ED X. In order to estimate the abandonment rate θ, we use techniques developed to analyze interval-censored data (see Chen et al. 2012) because we only observe the range within which the patience time of each patient lies rather than the exact values, similar to Batt and Terwiesch (2015). We assume that patients with the same ESI level have the same abandonment rate. Treatment bed capacity, M, is estimated from the average census of the patients in treatment beds during peak hours of the day for ESI Level 3 patients for simplicity. Because we mainly focus on the busiest time for the ED, we assume that all available beds are staffed and M is time-invariant for simplicity. This also allows us to focus on the impact of other parameters on the accuracy of the proposed method (although we assume that M is fixed throughout the day for simplicity, recall that our solution procedure is general and can accommodate time-variant capacity).

We use ED X data to estimate treatment times as well, but we are not able estimate the precise impact of PT on treatment times directly because ED X implemented PT most of the weekdays and the patient mix arriving at the ED on the weekends is significantly different from that on weekdays. Therefore, we use the number of tests ordered as a proxy to estimate the reduction in time spent in the treatment step if PT is applied. For the arrival rate, we use the number of ESI Level 3 patients arriving at the ED in each hour, which ranges from 1 to 5 patients/hour throughout the day. The parameter estimates are presented in Table 1.

Parameter	Value	Parameter	Value
r_2	\$488.25/patient	μ_{2N}	0.166 patients/hr
r_3	\$263.66/patient	μ_{2P}	0.221 patients/hr
w_{2N}	\$13.18/patient	μ_{3N}	0.181 patients/hr
w_{2P}	\$59.59/patient	θ	0.13 patients/hr
w_{3N}	\$0/patient	γ	0.18
c_N	\$32.66	M (nb. of servers)	19
c_P	\$124.04		

Table 1 Parameter estimates.

Simulation Details: We consider the patient flow in the ED for a 24-hour period with the arrival rates we estimated from ED X data. Using Proposition 1 and the estimated parameter values, we obtain the suggested policy. In practice, an ED can apply different triage methods throughout the day. However, it is impractical to switch too frequently as this requires updating

the working procedures for personnel as well as relocating staff from one area of the ED to another. For example, a provider working in the treatment area typically treats multiple patients gathered in this area and so cannot switch between the treatment and triage areas frequently because the triage rooms are typically located in a different area from the treatment beds. Therefore, we assume that the triage method is *fixed* for each hour. [1] We estimate the revenue under the suggested policy via simulation and assume that interarrival, service, and patience times are exponential.

In order to benchmark the performance of our policy, we compare its performance to a policy obtained using simulation and total enumeration. With a slight abuse of terminology, we refer to this policy as the "optimal" policy, although it is only optimal (ignoring the variability in simulation results) among policies that have fixed triage decisions in each hour. The optimal policy is obtained as follows. We assume that NT is applied between 12 a.m. and 8 a.m., when arrival rates are at their lowest, to reduce computational burden. Then we simulate each possible policy 100 times to estimate its expected performance and pick the best one with the highest value of the objective function. (Even after these simplifications, it takes well over 24 hours of computation to obtain the optimal policy.) The system is assumed to start empty at 12 a.m. for simplicity, and we include revenues from patients whose treatments had been initiated but were not completed at the end of the simulation for revenue calculations, which constitute only 0.3% of the total revenue. We also ran simulations to assess the robustness of our results to various assumptions; see Remark 5 below for details.

We consider eight different scenarios to verify the effectiveness of our method. The parameter values estimated using the procedure above (see Table 1) are taken as the base parameter set. This set is referred to as Scenario 1 in Tables 2–4. We obtain seven additional scenarios by changing the treatment rates and number of treatment beds and keeping all other parameters the same in order to assess the robustness of our solution; see Table 2 for details. We denote the threshold arrival rate for applying PT as Λ_1 in Table 2, where Λ_1 is computed as in (21) using Proposition 1 because Conditions 1–3 in (18)–(20) are satisfied for all scenarios.

Results and discussion: The time intervals when PT is applied in the ED in the optimal and the proposed policies are provided in Table 3. In Table 4 we compare the average value of the objective function under the optimal policy, the proposed policy, and two extreme policies when NT and PT are applied between 8 a.m. and 12 a.m. (recall that the triage method is fixed as

[1] If the triage physician is moved to the triage area from the treatment area, it might be possible to estimate the cost of switching the location of the physician. In Appendix **??**, we present an integer program based on our approximations in (15) that can be used to additionally account for the cost of switching to determine when to use PT.

Scenario	μ_{2N}	μ_{3N}	μ_{2P}	M	Λ_1
1	0.166	0.181	0.221	19	3.52
2	0.191	0.206	0.221	19	4.00
3	0.217	0.232	0.221	19	4.50
4	0.243	0.258	0.246	19	5.00
5	0.166	0.181	0.221	22	4.03
6	0.166	0.181	0.221	25	4.54
7	0.166	0.181	0.221	28	5.04
8	0.166	0.181	0.221	31	5.55

Table 2 Parameter sets for each scenario.

Scenario	Optimal policy PT hours	Proposed policy PT hours
1	8 a.m.-8 p.m.	10 a.m.-10 p.m.
2	9 a.m.-8 p.m.	10 a.m.-8 p.m.
3	12 p.m.-6 p.m., 7 p.m.-8 p.m.	11 a.m.-7 p.m.
4	12 p.m.-8 p.m.	12 p.m.-2 p.m.
5	9 a.m.-7 p.m.	10 a.m.-8 p.m.
6	10 a.m.-3 p.m.	11 a.m.-7 p.m.
7	12 p.m.-1 p.m.	12 p.m.-2 p.m.
8	-	-

Table 3 Optimal and proposed policies for each scenario.

Scenario	Proposed policy	Always NT	Always PT
1	0.82%	9.04%	0.49%
2	0.12%	5.51%	0.88%
3	0.28%	1.77%	1.32%
4	0.41%	0.56%	1.57%
5	0.15%	5.28%	1.10%
6	0.58%	1.45%	2.74%
7	0.18%	0.12%	4.06%
8	0	0	4.47%
Average (max) % difference	0.32% (0.82%)	2.97% (9.04%)	2.08% (4.47%)

Table 4 Percent difference in the objective function under the optimal policy and other benchmark policies.

NT between 12 a.m. and 8 a.m. for all policies). We also provide the percentage difference in the objective function under the proposed solution, always NT, and always PT policies versus that of the optimal policy. The comparisons reported in Table 4 are based on simulations with 10,000 replications. The maximum width of 95% confidence intervals of the mean ratio of the objective function under each policy to that under optimal policy is less than 0.002 in these simulations; hence, we did not include them in the table.

In general, we see that NT is replaced by PT earlier in the day in the optimal policy. This is because the proposed policy recommends PT only when the arrival rate in the current period is

sufficiently high. However, the optimal policy might recommend increasing capacity through PT as a proactive solution to an increase in future arrival rates to ensure availability of more treatment beds when the arrival rates become higher. (We also verified that starting PT an hour earlier from our suggested solution might improve performance slightly in our simulations.)

As seen in Table 4, our proposed solution methodology performs well under various parameter values (with different load factors on treatment step). Its performance is very close to those under the optimal policy, with an average difference of 0.32% and a median difference of 0.22% from the optimal policy contribution. In addition, our solution approach is able to capture the patterns in the optimal policy both when the treatment capacity is low, as in Scenario 1, and when it is fairly high, as in Scenario 8 (see Table 3). Because obtaining the optimal triage policy in an ED by enumerating 2^{24} different policies would take approximately four months using a standard PC, slight reduction of the objective function under the proposed policy, which can be determined instantly, should be acceptable.

Remark 5. We have verified the robustness of our findings to changes in certain assumptions in our simulations and observed that our findings were not affected significantly. Specifically, we ran the following robustness checks, (i) We changed the service and abandonment time distributions to log-normal with coefficient of variation equal to 0.25, 0.5, and 1, while keeping means the same. (ii) To check the dependence of our results on the initial state of the system, we have simulated the ED for 110 consecutive days and used the first 10 days as a warm-up period when we find the optimal solution. (iii) We made the patience times dependent on the triage method by increasing the mean patience times by first 25% and then 50% for patients who are triaged by a physician. In all these experiments, the proposed policy performs significantly better than both "Always NT" and "Always PT", and in the eight scenarios we considered, it performs within 0.5% of the optimal policy on average in each parameter setting. In addition, we conducted a sensitivity analysis for the threshold arrival rates of the proposed policy with respect to various model parameters. The details are presented in Appendix **??**.

6. Conclusions and Insights

In this paper, we propose an analytical method to determine when to use NT or PT in EDs. Under the more traditional NT method, patients are triaged only by a nurse. On the other hand, under the PT method, patients are additionally triaged by a provider; and thus, some patients may be discharged after triage, and for others diagnostic tests can be ordered at the triage stage. This potentially reduces the workload for the treatment stage, which is typically the bottleneck in ED

operations. However, the PT method can also increase the costs of operating the ED because of (i) potentially higher staffing costs due to additional physicans at the triage step, and (ii) potentially higher abandonment costs since the hospital may not be fully reimbursed for the more intensive care provided at the PT step (relative to the NT step) if patients abandon the ED.

In order to determine which triage method is more effective, we developed closed-form approximations for the system performance using a queueing model and steady-state many-server fluid approximations. We then used an objective function that consists of revenues from patients and differential costs under these methods to determine the optimal triage method. We demonstrated the application of this method in ED X and showed that NT outperforms PT when the arrival rate is sufficiently low. However, in general, PT is preferable as the arrival rate increases due to the fact that the ED capacity increases under PT. We also showed that NT may outperform PT once the arrival rate becomes sufficiently high.

Managerial Insights: Our study provides several insights for ED managers.

- We observe that triage method decisions are driven by revenue, cost, and treatment times associated with the two triage methods considered. Although PT reduces time spent in treatment beds and increases throughput, we find that its use is not always economically justified (such as when the additional staffing cost from applying PT is high and the increase in treatment rates under PT vs. NT is not significant). For instance, Figure 3a shows an example where NT outperforms PT no matter how high arrival rates are.
- We observe a significant benefit (up to 9% increase in the objective function) from using our proposed solution to make judicious triage method decisions. What is more important from a practical perspective is that the comparison between NT and PT methods can be made easily using the approximations we provide in this paper. Also, our method has a low informational burden because managers only need to know the cost, revenue, and treatment rates associated with each triage method and the arrival rate pattern in the ED. In addition, the triage method policy we recommend typically has a threshold structure based on the arrival rates and thus is easy to implement in practice. In the setting of ED X, for example, our proposed solution recommends changing triage methods at up to two times during the course of a typical weekday.
- We also find that the intuition that PT becomes increasingly advantageous over NT at higher arrival rates does not always hold, although generally true in our numerical examples. When abandonment cost per patient under PT is sufficiently higher than that under NT, higher arrival rates increase costs under PT beyond a certain point more so than under NT. Therefore, it is possible that NT may be preferred over PT again for sufficiently high arrival rates (see Figure 3c for such a case in ED X setting).

Limitations: Although we presented numerous extensions to our basic model, there are limitations of our approach that should be taken into account during implementation: (i) We mainly focus on abandonments, but waiting time measures are also commonly used for measuring performance in EDs. (ii) We use steady-state approximations to determine the system performance. This may not be very accurate for systems with highly varying arrival rates. Our simulations showed, however, that this did not present a problem for the arrival pattern we observed in ED X, which is very typical among most EDs. (iii) Our method relies on fluid approximations for large systems, and these approximations may not be very accurate for small systems. (iv) Our insights are based on the PT implementation in ED X, and they may not hold for EDs that have implemented significantly different versions of PT or for EDs that have factors we have not accounted for in our model. For example, we leave a detailed analysis of the impact of misclassifications or the impact of time-varying treatment step capacity on triage method choice for future research – although some basic insights are discussed based on our current model. In general, our method can be used instead of "intuition" to determine when to implement PT, and high-fidelity simulations can be used to determine the precise impact of PT on various performance measures of interest once triage method decision is fixed or narrowed down to a few choices.

References

ACEP. 2006. Approaching full capacity in the emergency department. *American College of Emergency Physicians* .

Allon, G., S. Deo, W. Lin. 2013. The impact of size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Operations Research* **61**(3) 544–562.

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

Atar, R., C. Giat, N. Shimkin. 2010. The cμ/θ rule for many-server queues with abandonment. *Operations Research* **58**(5) 1427–1439.

Bassamboo, A., J.M. Harrison, A. Zeevi. 2006. Design and control of a large call center: Asymptotic analysis of an LP-based method. *Operations Research* **54**(3) 419–435.

Batt, R.J., C. Terwiesch. 2015. Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science* **61**(1) 39–59.

Batt, R.J., C. Terwiesch. 2016. Early task initiation and other load-adaptive mechanisms in the Emergency Department. *Management Science* Forthcoming.

Besbes, O., C. Maglaras. 2009. Revenue optimization for a make-to-order queue in an uncertain market environment. *Operations Research* **57**(6) 1438–1450.

BLS. 2012. May 2012 national occupational employment and wage estimates in United States-Bureau of Labor Statistics. http://www.bls.gov/oes/2012/may/oes_nat.htm.

Bramson, M. 1998. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems* **30**(1-2) 89–140.

Burström, L., M. Nordberg, G. Örnung, M. Castrén, T. Wiklund, M.L. Engström, M. Enlund. 2012. Physician-led team triage based on lean principles may be superior for efficiency and quality? A comparison of three emergency departments with different triage models. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* **20**(1) 1–10.

Chan, T.C., J.P. Killeen, D. Kelly, D.A. Guss. 2005. Impact of rapid entry and accelerated care at triage on reducing emergency department patient wait times, lengths of stay, and rate of left without being seen. *Annals of Emergency Medicine* **46**(6) 491–497.

Chen, D.G.D., J. Sun, K.E. Peace. 2012. *Interval-censored time-to-event data: Methods and applications*. CRC Press, Boca Raton, FL.

Cheung, D.S., J.J. Kelly, C. Beach, R.P. Berkeley, R.A. Bitterman, R.I. Broida, W.C. Dalsey, H.L. Farley, D.C. Fuller, D.J. Garvey, et al. 2010. Improving handoffs in the emergency department. *Annals of Emergency Medicine* **55**(2) 171–180.

Choi, Y.F., T.W. Wong, C.C. Lau. 2006. Triage rapid initial assessment by doctor (TRIAD) improves waiting time and processing time of the emergency department. *Emergency Medicine Journal* **23**(4) 262–265.

Clark, C. 2010. Hospital leaders give strategies to remove ED bottlenecks. http://www.healthleadersmedia.com/page-2/LED-247042/Hospital-Leaders-Give-Strategies-to-Remove-ED-Bottlenecks.

Cooke, M.W., P. Arora, S. Mason. 2003. Discharge from triage: modelling the potential in different types of emergency department. *Emergency Medicine Journal* **20**(2) 131–133.

Dai, J.G., T. Tezcan. 2011. State space collapse in many-server diffusion limits of parallel server systems. *Mathematics of Operations Research* **36**(2) 271–320.

Dobson, G., T. Tezcan, V. Tilson. 2013. Optimal workflow decisions for investigators in systems with interruptions. *Management Science* **59**(5) 1125–1141.

GAO. 2009. Government Accountability Office. Hospital emergency departments: Crowding continues to occur, and some patients wait longer than recommended time frames. http://www.gao.gov/assets/290/289048.pdf.

Gilboy, N., T. Tanabe, D. Travers, A.M. Rosenau. 2011. Emergency Severity Index (ESI): A triage tool for emergency department care, version 4. Implementation handbook 2012 edition. http://www.ahrq.gov/sites/default/files/wysiwyg/professionals/systems/hospital/esi/esihandbk.pdf.

Green, L.V., P. Kolesar. 1991. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science* **37**(1) 84–97.

Green, L.V., P. Kolesar, A. Svoronos. 1991. Some effects of nonstationarity on multiserver markovian queueing systems. *Operations Research* **39**(3) 502–511.

Green, L.V., J. Soares, J.F. Giglio, R.A. Green. 2006. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* **13**(1) 61–68.

Han, J.H., D.J. France, S.R. Levin, I.D. Jones, A.B. Storrow, D. Aronsky. 2010. The effect of physician triage on emergency department length of stay. *The Journal of Emergency Medicine* **39**(2) 227–233.

He, Y. 2013. Patient flow interventions and prioritization in emergency department. Master's thesis, Pennsylvania State University.

Henneman, P.L., M. Lemanski, H.A. Smithline, A. Tomaszewski, J.A. Mayforth. 2009. Emergency department admissions are more profitable than non–emergency department admissions. *Annals of Emergency Medicine* **53**(2) 249–255.

Holm, L.B., F.A. Dahl. 2009. Simulating the effect of physician triage in the emergency department of Akershus University Hospital. *Winter Simulation Conference*. 1896–1905.

Holroyd, B.R., M.J. Bullard, K. Latoszek, D. Gordon, S. Allen, S. Tam, S. Blitz, P. Yoon, B.H Rowe. 2007. Impact of a triage liaison physician on emergency department overcrowding and throughput: A randomized controlled trial. *Academic Emergency Medicine* **14**(8) 702–708.

Huang, J., B. Carmeli, A. Mandelbaum. 2012. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research* **63**(4) 892–908.

Imperato, J., D.S. Morris, D. Binder, C. Fischer, J. Patrick, L.D. Sanchez, G. Setnik. 2012. Physician in triage improves emergency department patient throughput. *Internal and Emergency Medicine* **7**(5) 457–462.

Kc, D.S. 2013. Does multitasking improve performance? Evidence from the emergency department. *Manufacturing & Service Operations Management* **16**(2) 168–183.

McHugh, M., P. Tanabe, M. McClelland, R.K. Khare. 2012. More patients are triaged using the Emergency Severity Index than any other triage acuity system in the United States. *Academic Emergency Medicine* **19**(1) 106–109.

Olshaker, J.S., N.K. Rathlev. 2006. Emergency department overcrowding and ambulance diversion: The impact and potential solutions of extended boarding of admitted patients in the emergency department. *The Journal of Emergency Medicine* **30**(3) 351–356.

Oredsson, S., H. Jonsson, J. Rognes, L. Lind, K. E. Goransson, A. Ehrenberg, K. Asplund, M. Castrén, N. Farrohknia. 2011. A systematic review of triage-related interventions to improve patient flow in emergency departments. *Scandinavian Journal of Trauma Resuscitation and Emergency Medicine* **19**(1) 43.

Pines, J.M., S. Iyer, M. Disbot, J. E. Hollander, F.S. Shofer, E.M. Datner. 2008. The effect of emergency department crowding on patient satisfaction for admitted patients. *Academic Emergency Medicine* **15**(9) 825–831.

Pinker, E.J., T. Tezcan. 2013. Determining the optimal configuration of hospital inpatient rooms in the presence of isolation patients. *Operations Research* **61**(6) 1259–1276.

Pitts, S.R., J.M. Pines, M.T. Handrigan, A.L. Kellermann. 2012. National trends in emergency department occupancy, 2001 to 2008: Effect of inpatient admissions versus emergency department practice intensity. *Annals of Emergency Medicine* **60**(6) 679–686.

Rogers, T., N. Ross, D. Spooner. 2004. Evaluation of a 'See and Treat' pilot study introduced to an emergency department. *Accident and Emergency Nursing* **12**(1) 24–27.

Rogg, J. G., B.A. White, P.D. Biddinger, Y. Chang, D.F.M. Brown. 2013. A long-term analysis of physician triage screening in the emergency department. *Academic Emergency Medicine* **20**(4) 374–380.

Saghafian, S., G. Austin, S.J. Traub. 2015. Operations research/management contributions to emergency department patient flow optimization: Review and research prospects. *IIE Transactions on Healthcare Systems Engineering* **5**(2) 101–123.

Saghafian, S., W.J. Hopp, M.P. Van Oyen, J.S. Desmond, S.L. Kronick. 2012. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research* **60**(5) 1080–1097.

Saghafian, S., W.J. Hopp, M.P. Van Oyen, J.S. Desmond, S.L. Kronick. 2014. Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing and Service Operations Management* **16**(3) 329–345.

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

Shumsky, R.A., E.J. Pinker. 2003. Gatekeepers and referrals in services. *Management Science* **49**(7) 839–856.

Song, H., A.L. Tucker, K.L. Murrell. 2015. The diseconomies of queue pooling: An empirical investigation of Emergency Department length of stay. *Management Science* **61**(12) 3032–3053.

Soremekun, O.A., P.D. Biddinger, B.A. White, J.R. Sinclair, S.B. Chang, Y.and Carignan, D.F.M. Brown. 2012. Operational and financial impact of physician screening in the ED. *The American Journal of Emergency Medicine* **30**(4) 532–539.

Stolyar, A.L., T. Tezcan. 2011. Shadow-routing based control of flexible multiserver pools in overload. *Operations Research* **59**(6) 1427–1444.

Subash, F., F. Dunn, B. McNicholl, J. Marlow. 2004. Team triage improves emergency department efficiency. *Emergency Medicine Journal* **21**(5) 542–544.

Terris, J., P. Leman, N. O'Connor, R. Wood. 2004. Making an IMPACT on emergency department flow: Improving patient processing assisted by consultant at triage. *Emergency Medicine Journal* **21**(5) 537–541.

Traub, S.J., J.P. Wood, J. Kelley, D.M. Nestler, Y.H. Chang, S. Saghafian, C.A. Lipinski. 2015. Emergency department rapid medical assessment: Overall effect and mechanistic considerations. *The Journal of emergency medicine* **48**(5) 620–627.

Travers, J.P., F.C.Y. Lee. 2006. Avoiding prolonged waiting time during busy periods in the emergency department: is there a role for the senior emergency physician in triage? *European Journal of Emergency Medicine* **13**(6) 342–348.

Whitt, W. 2004. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science* **50**(10) 1449–1461.

Wiler, J.L., C. Gentle, J.M. Halfpenny, A. Heins, A. Mehrotra, M.G. Mikhail, D. Fite. 2010. Optimizing emergency department front-end operations. *Annals of Emergency Medicine* **55**(2) 142–160.

Xu, K., C.W. Chan. 2016. Using future information to reduce waiting times in the emergency department via diversion. *Manufacturing & Service Operations Management* **18**(3) 314–331.

Ye, K., D. McD Taylor, J.C. Knott, A. Dent, C.E. MacBean. 2007. Handover in the emergency department: Deficiencies and adverse effects. *Emergency Medicine Australasia* **19**(5) 433–441.

Yom-Tov, G.B., A. Mandelbaum. 2014. Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management* **16**(2) 283–299.

Zayas-Cabán, G., J. Xie, L.V. Green, M.E. Lewis. 2014. Optimal control of an Emergency Room triage and treatment process Working Paper No. 14-51, Columbia Business School, New York.