When to Use Provider Triage in Emergency Departments

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 (Pitts et al. (2012), GAO (2009)). 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), Mohsin et al. (2007) and Batt and Terwiesch (2015)). The U.S. Government Accountability Office 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 treatment 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, 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, patient is seen by a provider for the first time at the triage stage instead of at 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 (Choi et al. (2006), Subash et al. (2004), Terris et al. (2004), Travers and Lee (2006)). Patients with more severe conditions, however, are sent to a waiting room once their triage interventions are completed where they wait for an available bed in the treatment area. Again, patients with certain conditions may follow a different route. 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 at one ED that we worked with (referred to as ED X throughout this paper), but the patient flow and treatment decisions in other EDs in the U.S. are also very similar (see Rogg et al. (2013) and Soremekun et al. (2012)). Each patient who arrives at 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)). The most severe cases (ESI Level 1) and the least severe cases (ESI Levels 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). As such, patients with ESI Levels 2 and 3 are the most severely affected by ED crowding because they have to wait for a treatment bed before being treated. In this paper, we mainly consider the triage method decisions regarding patients with ESI Levels 2 and 3 in our model.
PT affects the patient flow and 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 the 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 (Han et al. (2010), Holroyd et al. (2007), Subash et al. (2004)). 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. Thus, when PT is applied, fewer patients end up needing treatment beds, and those who are assigned treatment beds spend less time there. Thus, EDs’ treatment capacity per bed will increase if PT is applied. 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 staffing level in the ED. 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 a triage method during the course of a day in EDs. The general practice is to deploy PT when the patient volume is “high” (see for example Han et al. (2010) and Holroyd et al. (2007)). 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 PT practice is sensible. To achieve this we develop a queueing model that captures 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) and Whitt (2006)). 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—the ED of a teaching hospital that 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 in terms of profit (a $0.32\%$ reduction in profits 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.

## 2. Literature Review

There are several comprehensive reviews of the alternative triage system implementations in the medical literature (see Gilboy et al. (2011), and Moskop and Iserson (2007) among others). Alternative triage methods have also been studied in the operations management literature, (see Saghaian 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), Day et al. (2013), 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 and 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 a clue 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.

Our study is closely related to the branch of literature that analyzes patient flows in EDs such as Huang et al. (2012), Dobson et al. (2013), Saghafian et al. (2012). In order to capture the operational and economic aspects of PT vs. NT, we model the ED as a two-step service system where the first step is triage and the second step is treatment, similar to Zayas-Cabán et al. (2014). In Zayas-Cabán et al. (2014), the two steps are carried out by the same provider, and the question is how to prioritize the work of the provider based on delays. However, in our model, the providers in the two steps, if one is involved in the first step, are different and priorities are defined by the acuity of the patients based on medical regulations. While they take the provider as given in the initial step of the process, we take whether to involve a provider in the first step as our decision variable. 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
(IP) patients). But 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 discuss the different ways PT is implemented in practice in §3.1. We then present our queueing models 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 the way PT is implemented can vary. For notational and expositional simplicity, we present our solution methodology based on the implementation in ED X. However, our model is flexible enough to incorporate the documented differences in practice. 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 patients with minor complaints and discharge them from the PT area, and initiate the treatment of more severe patients and place them in the queue for a treatment bed (ACEP (2006), Chan et al. (2005), Traub et al. (2016)). 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 “PT for Severe Patients”, 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 Levels 2 and 3, and only ESI Levels 2 and 3 patients are treated by a provider in triage. This will be our base model. In other implementations of PT, provider at triage can also treat ESI Levels 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 NT step and then additionally triaged by a provider at the PT step. Finally, they undergo the treatment step.

We divide patients into two groups, triage or treatment, based on the resources they require for treatment completion. Triage 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. Treatment 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 treatment patients. ESI Level 3 patients are less urgent than ESI Level 2 patients and can be either triage or treatment patients, e.g., in ED X 18% of ESI Level 3 patients are triage patients. (The classification of patients as triage or treatment 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 patient registration. 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 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.

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 treatment patients have higher (non-preemptive) priority than ESI Level 3 triage patients. 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.

**Figure 1** Patient flow when the NT method is applied.

*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 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, triage patients are discharged after the PT step, and all other patients are placed in queue for a treatment bed, where ESI Level 2 patients are given (non-preemptive) priority over ESI Level 3 treatment patients.

### 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 treatment patients as Type 2 patients, and ESI Level 3 triage 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 \( M_j(t) \) be 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 \( \delta_i \) and \( \mu_{ij} \), 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 – e.g., 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 of abandonment. 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 –84.19% in ED X– arrive at the ED. Let $\pi$ denote a triage policy, and for notational simplicity we also define $\pi$ 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 profit function consists of three components. The first part is the revenue earned from patients treated in the ED. We define $S_{ij}^\pi(t)$ and $D_{ij}^\pi(t)$ 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 $B_{ij}^\pi(t)$ and $E_{ij}^\pi(t)$, 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

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

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

$$\Phi^\pi(T) = \sum_{i=1}^3 \sum_{j \in \{N, P\}} r_{ij} (E[D_{ij}^\pi(T)] + E[S_{ij}^\pi(T)]) - \sum_{i=1}^3 \sum_{j \in \{N, P\}} (y_i E[B_{ij}^\pi(T)] + w_{ij} E[E_{ij}^\pi(T)]) - K^\pi(T).$$

1 We consider an alternative objective where the goal is to minimize the number of abandonments in §4.3.
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 $E[D_{ij}(T)]$, $E[E_{ij}(T)]$, $E[B_{ij}(T)]$ and $E[S_{ij}(T)]$ 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 insights, we use fluid approximations. We explain this 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 $\lambda$, fraction of arrivals $\gamma_i$ and arrival rate $\lambda_i$ 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

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

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

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

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

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

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

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

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

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

\[
b_{iP}(\lambda) = \gamma_i \left( \lambda - \left( \sum_{i=1}^{3} \frac{\gamma_i}{\delta_i} \right)^{-1} \right)^+ , \quad i = 1, 2, 3,
\]

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

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

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

\[
\kappa_{iP}(\lambda) = \lambda_i - b_{iP}(\lambda) - s_{iP}(\lambda), \quad i = 1, 2, 3.
\]

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

\[
M_{1P}(\lambda, M_P) = \min \left\{ M_P, \frac{\kappa_{1P}(\lambda)}{\mu_{1P}} \right\}, \quad M_{2P}(\lambda, M_P) = \min \left\{ M_P - M_{1P}(\lambda, M_P), \frac{\kappa_{2P}(\lambda)}{\mu_{2P}} \right\}, \quad M_{3P}(\lambda, M_P) = 0.
\]
Thus, the rate at which type $i$ patients are discharged from the treatment area is
\[
d_{iP}(\lambda, M_P) = \mu_{iP} M_{iP}(\lambda, M_P),
\] (13)
and the rate at which type $i$ patients abandon the treatment bed queue is
\[
a_{iP}(\lambda, M_P) = \kappa_{iP}(\lambda) - \mu_{iP} M_{iP}(\lambda, M_P), \quad i = 1, 2, 3.
\] (14)

**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
\[
b_{iP}(\lambda) = (\lambda_1 - \delta_1)^+, \quad b_{2P}(\lambda) = \left(\lambda_2 - \left(1 - \frac{\lambda_1}{\delta_1}\right)^+ \delta_2 \right)^+, \quad b_{3P}(\lambda) = \left(\lambda_3 - \left(1 - \frac{\lambda_1}{\delta_1} - \frac{\lambda_2}{\delta_2}\right)^+ \delta_3 \right)^+.
\] (15)

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

**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 $\lambda$, 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;
\[
\Theta^N(\lambda, M_N) = \sum_{i=1}^{3} r_i N d_{iN}(\lambda, M_N) - \sum_{i=1}^{3} w_i N a_{iN}(\lambda, M_N) - c_N,
\]
\[
\Theta^P(\lambda, M_P) = \sum_{i=1}^{3} r_i P d_{iP}(\lambda, M_P) + s_{iP}(\lambda) - \sum_{i=1}^{3} (y_i b_{iP}(\lambda) + w_i P a_{iP}(\lambda, M_P)) - c_P.
\] (16)

Using (16), we arrive at the following approximation $\hat{\Phi}^\pi(T)$ for profit function $\Phi^\pi$ under policy $\pi$
\[
\hat{\Phi}^\pi(T) = \left[ \int_0^T \Theta^{\pi(t)}(\lambda(t), M_{\pi(t)}(t)) dt \right].
\] (17)
One of the important features of our approximations is that because they are based on steady-state quantities, the approximation at each time point is independent of the state of the system prior to that point. 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{18}

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

4.3. Extensions

We consider several additional aspects of PT as extensions of our base model for which the details are presented in Appendix A. We give a summary here. i) In certain applications of PT, ESI Levels 4 and 5 patients are also triaged by the provider. We present the extension of our model to this case in Appendix A.1. 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 A.2. We also explain how “See and Treat” and “PT for Severe Patients” implementations can be analyzed using this extension. iii) The misclassification of patients in triage has been documented in the literature (see for example Saghafian et al. (2014)). We extend our model to account for misclassifications in Appendix A.3. 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 A.4, we provide an alternate objective function of minimizing the number of abandonments. v) In Appendix A.5 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, i.e., when treatment and triage 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 accuracy 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 F.) Hence, the staffing cost is higher under the PT method, i.e., $c_P > c_N$.

Based on the implementation details in ED X, we make two simplifying assumptions in our model. First, we only consider one type of triage patients, and we 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 off 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 insights on the triage decisions. Even when these assumptions do not hold exactly, the insights 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 the next section).

We also make the following additional assumptions. We assume that the number of (staffed) treatment beds is the same under both triage methods, i.e., $M = M_N = M_P$. We also observed that patients rarely wait before the PT step. This is because patients are placed in the waiting room after the brief evaluation by the triage provider and do not occupy resources 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 treatment patients is higher than that for the triage patients, i.e., $w_{2j} \geq w_{3j}$ 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 $\mu_{2P} > \mu_{2N}$, i.e., the treatment patients are treated more quickly under PT than NT (see Appendix D for the analysis of the case with $\mu_{2P} \leq \mu_{2N}$). Because treatment procedures for treatment patients are more complex 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 $\gamma$, 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 \( \lambda \) 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 \( \Theta^N \) and \( \Theta^P \));

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 \( \lambda \leq \Lambda_2 \) and \( j^*(\lambda, M) = P \) for all \( \lambda > \Lambda_2 \). 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 \( \Theta^N \) and \( \Theta^P \) 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.

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)
   \[
   \Lambda_1 = \frac{\mu_{2N}(M\mu_{3N}(r_3 + w_{3N}) + c_P - c_N)}{(r_3 + w_{3N})(\gamma\mu_{2N} + (1-\gamma)\mu_{3N})},
   \]
   
   if Condition 1 holds,

   (b)
   \[
   \Lambda_1 = \frac{(r_2 + w_{2N})M\mu_{2N} + c_P - c_N}{\gamma(r_3 + w_{3N}) + (1-\gamma)(r_2 + w_{2N})},
   \]
   
   if Condition 1 does not hold but Condition 2 holds,
\[ \Lambda_1 = \frac{(r_2 + w_{2N})M\mu_{2N} - (r_2 + w_{2P})M\mu_{2P} + c_P - c_N}{\gamma(r_3 + w_{3N}) - (1 - \gamma)(w_{2P} - w_{2N})}, \]  
\[ (24) \]

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) \( \Lambda_1 \) is given by (22), and \( \Lambda_2 \) is given by the right-hand side of (24) if Condition 1 holds,
(b) \( \Lambda_1 \) is given by (23), and \( \Lambda_2 \) is given by the right-hand side of (24) 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 B.

![Graph](image)

(a) No Intersection.

![Graph](image)

(b) One Intersection.

![Graph](image)

(c) Two Intersections.

**Figure 3** \( \Theta^N(\lambda, M) \) and \( \Theta^P(\lambda, M) \) in ED X vs. \( \lambda \) with zero, one or two intersections.
Sketch of the proof of Proposition 1: Figure 3 illustrates how the profit functions $\Theta^N$ and $\Theta^P$ 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 profits for both systems increase until the systems become overloaded. Because the system under NT has a lower capacity, it reaches full capacity (at point $\lambda_0$ in Figures 3(a)–(c)) before the system under PT does (at $\lambda_2$ in Figures 3(a)–(c)).

Once each system becomes overloaded, some of the patients will abandon the system. For the system under NT, initially triage patients will abandon after the arrival rate exceeds $\lambda_0$ because treatment patients have priority. When the arrival rate exceeds $\lambda_1$, some of the treatment patients will abandon as well because there is insufficient capacity to serve all the treatment patients. Hence, the slope of $\Theta^N$ changes at $\lambda_0$ and then again at $\lambda_1$. On the other hand, the slope of $\Theta^P$ changes only once at $\lambda_2$ because all the triage patients are treated at the triage step.

Condition 1 implies that $\Theta^P$ is greater than or equal to $\Theta^N$ at arrival rate $\lambda_1$. At $\lambda_1$, all patients can be treated if PT is used, whereas only treatment 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 $\lambda$. Similarly, Condition 2 implies that $\Theta^P$ is greater than or equal to $\Theta^N$ at $\lambda_2$, when all patients are treated if PT is used, but all of the triage and some of the treatment 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 $\Theta^P$ is greater than or equal to that of $\Theta^N$ beyond $\lambda_2$.

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

$$R^j(\lambda, M) = \sum_{i=2}^{3} r_i d_{ij}(\lambda, M), \ A^j(\lambda, M) = \sum_{i=2}^{3} w_{ij} a_{ij}(\lambda, M) \quad \text{for} \ j = N, P.$$
Figure 4 Abandonment cost and revenue functions under NT and PT.

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

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, ii) when a provider is moved to triage from the treatment step (see Appendix F for details) and iii) for the alternative implementations of PT discussed in §4.3 (see Appendix C). We next explain the case when PT has limited capacity. Consider the cost and the revenue functions in Figure 4(b) for this case (see Appendix E for a detailed analysis). If the total capacity of the PT step is $\lambda'_2$ 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 high. 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, such as abandonment costs and PT capacity depending on the implementation, that need to be considered, 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 Besbes and Maglaras (2009), Pinker and Tezcan (2013), Stolyar and Tezcan (2011), Bassamboo et al. (2006) 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 insights 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 \( T^n = k^nT \). Assume that the arrival rate \( \Lambda^n \) in the \( n^{th} \) ED satisfies

\[
\Lambda^n(t) = n\Lambda\left(\frac{t}{k^n}\right),
\]

for a nonnegative continuous function \( \Lambda \) satisfying

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

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

\[
M^n(t) = nM\left(\frac{t}{k^n}\right),
\]

for a nonnegative continuous function \( M \) satisfying

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

for some constant \( M_c < \infty \). Following (27), we assume that the staffing costs under NT and PT in \( n^{th} \) ED have the following form:

\[
c^n_j = nc_j, \quad j = N, P.
\]

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 \( \pi(t) = 1 \{ j^*(\Lambda(t), M(t)) = N \} \) has finitely many discontinuities in \([0,T]\).
First, we consider the solution for (2) in the $n$th ED. Let $\Pi$ denote the set of admissible policies. We set

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

Using (17), we also define

$$\Phi^* = \int_0^T \Theta^{\pi^*(t)}(\lambda(t), M(t)) dt, \quad (30)$$

where

$$\pi^*(t) = j^* (\Lambda(t), M(t)), \quad (31)$$

and $j^*$ is defined in (18). Also, our solution procedure calls for the use of triage policy $\pi^{*,n}(t)$ at time $t$, which is defined as follows:

$$\pi^{*,n}(t) = j^* (\Lambda^n(t), M^n(t)). \quad (32)$$

Note that by (25), (27) and (32), the arrival rates at which the suggested triage method changes under the threshold policy in the $n$th system are $n\Lambda_1$ and $n\Lambda_2$, where $\Lambda_1$ and $\Lambda_2$ are defined in Proposition 1.

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

$$\lim_{n \to \infty} \frac{Q^n(0)}{n} = Q(0) < \infty \quad \text{a.s.} \quad (33)$$

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 (25)--(29) and (33). Then,

$$\liminf_{n \to \infty} \Phi^{\pi^n}(T^n) \geq \Phi^* \quad (34)$$

under any sequence $\pi^n$ of admissible policies.

(ii) In addition, if Assumption 1 holds, then

$$\lim_{n \to \infty} \Phi^{\pi^{*,n}}(T^n) = \Phi^*. \quad (35)$$

Theorem 1 implies that if the triage method is selected according to (16) and (18), then the properly scaled profit of the triage policy is optimized asymptotically as the system becomes larger. In addition to providing the optimal triage method, (16) and (18) also provide asymptotically correct estimates of the profit under each threshold policy by (35).
Remark 4. The motivation for using the scaling (25) comes from empirical studies in the literature (Shi et al. (2015), Armony et al. (2015), Green et al. (2006), Yom-Tov and Mandelbaum (2014), Saghafian et al. (2012), Saghafian et al. (2014)) as well as our own observation that although the arrival rate in EDs is time-dependent, its value does not change significantly 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 at 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, i.e., 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 appropriate, mostly for cost and revenue parameters. However, most of the operational parameters that we use ($\mu_{ij}$, $\lambda$, $\gamma$, $M$ and $\theta$) are not reported in the literature; therefore, we use data obtained from ED X.

We use the values in Soremekun et al. (2012) and Henneman et al. (2009) to estimate the revenue per patient. We set the abandonment cost per patient as 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 again use the data from ED X. In order to estimate the abandonment rate $\theta$, we use techniques developed to analyze interval-censored data (see Chen et al. (2012) and Fay and Shaw (2010) for more on interval censored data. We implicitly assume that patients with the same ESI level have the same patience time distribution. In ED X data, 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). 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. The data revealed that ED X implemented PT during daytime on weekends. In addition, the patient mix arriving at the ED is different on weekdays compared to weekends. Hence, we cannot directly compare the impact of PT on treatment times. 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.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_2$</td>
<td>$488.25/$patient</td>
<td>$\mu_{2N}$</td>
<td>0.166 patients/hr</td>
</tr>
<tr>
<td>$r_3$</td>
<td>$263.66/$patient</td>
<td>$\mu_{2P}$</td>
<td>0.221 patients/hr</td>
</tr>
<tr>
<td>$w_{2N}$</td>
<td>$13.18/$patient</td>
<td>$\mu_{3N}$</td>
<td>0.181 patients/hr</td>
</tr>
<tr>
<td>$w_{2P}$</td>
<td>$59.59/$patient</td>
<td>$\theta$</td>
<td>0.13 patients/hr</td>
</tr>
<tr>
<td>$w_{3N}$</td>
<td>$0/$patient</td>
<td>$\gamma$</td>
<td>0.18</td>
</tr>
<tr>
<td>$c_N$</td>
<td>$32.66$</td>
<td>$M$ (nb. of servers)</td>
<td>19</td>
</tr>
<tr>
<td>$c_P$</td>
<td>$124.04$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1 Parameter estimates.

Testing the performance of the proposed solution methodology: 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. That provider cannot switch between the treatment and triage areas frequently because the triage rooms are in a different area from the treatment beds. Therefore, we assume that the triage method is fixed for each hour.

We estimate the revenue under the suggested policy via simulation and assume that interarrival, service and patience times are exponential. We compare this policy to the 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. (We use common random numbers to reduce variation.) To reduce computational burden, we assume that NT is applied between 12 a.m. and 8 a.m., when arrival rates are at their lowest. Even after this simplification, it takes over 24 hours of computation to find the optimal policy. We use 100 replications in each simulation, and the system is assumed to start empty at 12 a.m. Also, for profit calculations, we include revenues from
patients whose treatments had been initiated but were not completed at the end of the simulation, which constitute only 0.3% of the total profit.

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 \( \Lambda_1 \) in Table 2, where \( \Lambda_1 \) is computed as in (22) using Proposition 1 because Conditions 1–3 in (19)–(21) are satisfied for all scenarios.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>( \mu_{2N} )</th>
<th>( \mu_{3N} )</th>
<th>( \mu_{2P} )</th>
<th>( M )</th>
<th>( \Lambda_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.166</td>
<td>0.181</td>
<td>0.221</td>
<td>19</td>
<td>3.52</td>
</tr>
<tr>
<td>2</td>
<td>0.191</td>
<td>0.206</td>
<td>0.221</td>
<td>19</td>
<td>4.00</td>
</tr>
<tr>
<td>3</td>
<td>0.217</td>
<td>0.232</td>
<td>0.221</td>
<td>19</td>
<td>4.50</td>
</tr>
<tr>
<td>4</td>
<td>0.243</td>
<td>0.258</td>
<td>0.246</td>
<td>19</td>
<td>5.00</td>
</tr>
<tr>
<td>5</td>
<td>0.166</td>
<td>0.181</td>
<td>0.221</td>
<td>22</td>
<td>4.03</td>
</tr>
<tr>
<td>6</td>
<td>0.166</td>
<td>0.181</td>
<td>0.221</td>
<td>25</td>
<td>4.54</td>
</tr>
<tr>
<td>7</td>
<td>0.166</td>
<td>0.181</td>
<td>0.221</td>
<td>28</td>
<td>5.04</td>
</tr>
<tr>
<td>8</td>
<td>0.166</td>
<td>0.181</td>
<td>0.221</td>
<td>31</td>
<td>5.55</td>
</tr>
</tbody>
</table>

Table 2  Parameter sets for each scenario.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Optimal policy PT hours</th>
<th>Proposed policy PT hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8 a.m.-8 p.m.</td>
<td>10 a.m.-10 p.m.</td>
</tr>
<tr>
<td>2</td>
<td>9 a.m.-8 p.m.</td>
<td>10 a.m.-8 p.m.</td>
</tr>
<tr>
<td>3</td>
<td>12 p.m.-6 p.m., 7 p.m.-8 p.m.</td>
<td>11 a.m.-7 p.m.</td>
</tr>
<tr>
<td>4</td>
<td>12 p.m.-8 p.m.</td>
<td>12 p.m.-2 p.m.</td>
</tr>
<tr>
<td>5</td>
<td>9 a.m.-7 p.m.</td>
<td>10 a.m.-8 p.m.</td>
</tr>
<tr>
<td>6</td>
<td>10 a.m.-3 p.m.</td>
<td>11 a.m.-7 p.m.</td>
</tr>
<tr>
<td>7</td>
<td>12 p.m.-1 p.m.</td>
<td>12 p.m.-2 p.m.</td>
</tr>
<tr>
<td>8</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3  Optimal and proposed policies for each scenario.

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 profit for 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 NT between 12 a.m. and 8 a.m. for all policies). We also provide the percentage difference in profits of the proposed solution, always NT, and always PT policies vs. that of the optimal policy.
Table 4  % difference in profits of the optimal policy and other benchmark policies.

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

As seen in Table 4, our proposed solution methodology yields profits that are very close to those under the optimal policy under various parameter values, with an average difference of 0.32% and a median difference of 0.22% from the optimal policy contribution. The proposed approach performs well under various load factors in treatment rooms. A comparison of the proposed and optimal policies shows that 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. (We also obtained the 95% confidence intervals (CIs) of the mean ratio of the profit of each policy to the optimal profit by running 10,000 replications. The maximum width of the CI across eight scenarios is less than 0.002.)

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 increases sufficiently. However, the optimal policy might recommend increasing capacity through PT as a proactive solution to the future increase in future arrival rates and ensure the availability of more treatment beds when the arrival rates are high. Despite this difference, we see that the proposed and optimal policies have quite similar structures. Also, obtaining the optimal triage policy in the ED by enumerating $2^{24}$ different policies would take approximately four months using a standard PC, whereas the proposed policy can be obtained instantaneously. Thus, we conclude that our approach is able to yield near-optimal solutions using back-of-the-envelope calculations in this setting.

Sensitivity analysis: We next conduct a sensitivity analysis for the threshold arrival rates of the proposed policy with respect to various model parameters. We are especially interested in the sensitivity of the threshold values to the abandonment costs because they are arguably the most difficult parameters in our model to estimate. It is not easy to conduct a general sensitivity analysis because both the threshold arrival rates and the structure of the policy might change when
parameter values change. Therefore, we mainly focus on the case when the structure of the policy remains the same, i.e., the same part of Proposition 1 is applicable to all parameter combinations. We comment on the more general case at the end.

We first report the results for Scenario 1 specified in Table 2; the results for Scenario 8 are very similar (these two scenarios have the lowest and highest $\Lambda_1$ values). We examine the change in $\Lambda_1$ when we change one parameter at a time for three different values (0, 100 and 500) of the abandonment cost for type 3 patients under NT, $w_{3N}$. In all parameter combinations, the optimal policy is given in Part 2(a) of Proposition 1. (We note that this is true in general when PT is more profitable for relatively lower arrival rates; see Figure 3 and the sketch of the proof of Proposition 1 on page 18.) We present the plots of $\Lambda_1$ vs. $\mu_{2N}$, $\mu_{3N}$, $\gamma$ and $M$ when each ranges within $\pm 25\%$ from their base values in Figure 7 in Appendix H.

We observe that $\Lambda_1$ is most sensitive to $\mu_{2N}$ and $M$ but much less so to $w_{3N}$. Specifically, a 25% change in $M$ or $\mu_{2N}$ leads to as much as a 42% change in $\Lambda_1$ while increasing $w_{3N}$ from 0 to 500 leads to at most a 7.8% decrease in $\Lambda_1$ (and to less than a 4% decrease when it is increased from 0 to 100). Similarly, $\Lambda_1$ is not very sensitive to changes in $\mu_{3N}$ and $\gamma$; the maximum change we observed in $\Lambda_1$ is 4.8%.

To understand how $\Lambda_1$ in Parts 2(b) and 2(c) of Proposition 1 is affected by changes in abandonment costs, we conduct a similar sensitivity analysis. We use the parameter values in Case 1, except this time we first set $c_p = 500$ so that conditions of Part 2(b) of Proposition 1 are satisfied, and then set $c_p = 1000$ so that conditions of Part 2(c) of Proposition 1 are satisfied. We test the sensitivity of $\Lambda_1$ with respect to only $w_{3N}$ first, and then with respect to $w_{2N}$, $w_{2P}$ and $w_{3N}$ by simultaneously increasing all three values from the base value of 0. For $\Lambda_1$ in Part 2(b) of Proposition 1, the range of values we consider for $w_{3N}$ is between 0 and 400, beyond which conditions of Part 2(b) of Proposition 1 are no longer satisfied, and between 0 and 1000 for $w_{2N}$ and $w_{2P}$. In each setting, we assume $w_{2N} = w_{2P}$ for simplicity. We tested 10 combinations by increasing each one equally from one experiment to another so that they would reach their maximum values in the last experiment. The maximum change we observe in $\Lambda_1$ is a 13% decrease. For Part 2(c) of Proposition 1, the maximum value we tested for $w_{3N}$ was 200 (for the same reason stated above) and we observed a maximum change of 22% in $\Lambda_1$.

In conclusion, the threshold values in the proposed policy are relatively robust to estimation errors in abandonment costs (with the parameter values we considered). It is also possible to find analytical bounds for $\Lambda_1$ when Condition 1 or 2 holds, which then can be used to gain additional
insights in the sensitivity analysis. First, $\lambda_0 = M\mu_{avg}$, where $\mu_{avg} = [(1 - \gamma)/\mu_{2N} + \gamma/\mu_{3N}]^{-1}$, $\lambda_1 = M\mu_{2N}/(1 - \gamma)$ and $\lambda_2 = M\mu_{2P}/(1 - \gamma)$ (see Figure 3(b)), and therefore

$$\frac{\lambda_1}{\lambda_0} = 1 + \frac{\gamma - \mu_{2N}}{1 - \gamma \mu_{3N}} \text{ and } \frac{\lambda_2}{\lambda_1} = \frac{\mu_{2P}}{\mu_{2N}}.$$ (36)

Also, if Condition 1 holds then $\Lambda_1 \in [\lambda_0, \lambda_1]$, and if Condition 2 holds but Condition 1 does not, then $\Lambda_1 \in [\lambda_1, \lambda_2]$. Clearly, the values of $\lambda_0$, $\lambda_1$ and $\lambda_2$ do not depend on the abandonment costs. A change in abandonment costs might affect which conditions hold, but as long as Condition 1 or 2 is satisfied, $\Lambda_1 \in [\lambda_0, \lambda_2]$. In Scenario 1, $\lambda_1/\lambda_0 = 1.18$ and $\lambda_2/\lambda_1 = 1.33$; hence, $\Lambda_1$ is less sensitive to changes in the abandonment costs when Condition 1 or 2 holds.

So far we have made the simplifying assumption that the same conditions (among Conditions 1–3) hold even when the parameters change. The main reason is that Condition 1 is always satisfied with the parameter values we considered; and it does not hold only when we use impractical values for $c_P$. In general, we observe that for Condition 3 to hold, NT must be preferred to PT even at full capacity of PT, i.e., $\lambda = \lambda_2$, in addition to certain conditions on the abandonment costs. In this case, abandonment costs can potentially have more impact on $\Lambda_1$. A similar analysis can be carried out if conditions in Part 3 of Proposition 1 hold since the thresholds in that part are the same as those in Part 2.

6. Conclusion

In this paper, we compare two triage methods commonly used in EDs. In the provider triage (PT), patients are triaged by a provider in addition to a nurse. In the traditional nurse triage (NT) method, patients are triaged by a nurse only. Under PT, the triage provider initiates the medical assessment and orders diagnostic tests if necessary at the triage stage. This potentially reduces the workload for providers in the treatment area. Additionally, under PT, patients with minor conditions can be treated by the triage provider and discharged without occupying a treatment bed. On the downside, PT can potentially lead to an increase in the abandonment costs per patient since more tests are ordered during triage, and can affect the staffing cost. PT can also lead to a reduction in the treatment area capacity if the provider is moved to triage from the treatment area, or an increase in staffing cost if an additional provider is deployed in triage.

We formulate and solve the triage method optimization problem faced by EDs by considering the profits of the triage methods using a queueing framework. Through a steady-state many-server fluid approximation, we present closed-form expressions to approximate the profits under the two triage methods. Then, for the implementation in ED X, we obtain a closed-form expression for the range
of arrival rates for which PT outperforms NT economically. Our findings show that NT provides higher profits when the arrival rate is sufficiently low. As the arrival rate increases, higher revenues thanks to increased treatment capacity compensate for higher staffing costs and potentially higher abandonment costs during the hours when PT is used. Depending on the parameter values, however, NT may outperform PT economically once the arrival rate becomes sufficiently high. Based on parameter values estimated from the literature and ED X data, we demonstrate the accuracy of our suggested method under various scenarios.

To the best of our knowledge, our paper is the first to provide a relatively simple analytical method to determine whether and under what conditions PT is economically more beneficial than NT. Our results can be used in the first step of a PT implementation to verify whether an ED can benefit from switching to this alternative triage method. Although various studies in the extant literature have compared the operational outcomes in EDs after implementing PT, the decision on when PT should be implemented was based on crude intuition. In addition, because the environment in which the EDs operate changes constantly, these comparisons can potentially suffer from other confounding factors. However, fluid analysis in general does not provide accurate waiting-time estimates. These estimates can and should be assessed using simulations once PT is verified to be economically feasible.

References


Clark, C. 2010. Hospital leaders give strategies to remove ED bottlenecks. 


GAO. 2009. Government Accountability Office. Hospital emergency departments: Crowding continues to occur, and some patients wait longer than recommended time frames. 


A. Extensions of the base model

We next consider several extensions of our base model. We present the extension of our model to the case with ESI Levels 4 and 5 patients treated in PT in §A.1. In §A.2 we discuss how to use our approach when not necessarily all patients are seen by the triage provider under PT method. In §A.3 we explain how to incorporate misclassification of patients into our solution approach. We discuss an alternate objective function of minimizing abandonments from the ED in §A.4. Finally, §A.5 discusses the use of our solution methodology in EDs that prioritize patients only according to their acuity level (i.e., ESI level) in treatment bed queue.

A.1. ESI Levels 4 and 5 patients treated in PT

Our model can also be used when ESI Levels 4 and 5 patients go through PT step and share the main treatment area resources with ESI Levels 2 and 3 patients. For simplicity, consider ESI Levels 4 and 5 patients as a single patient class and refer to them as Type 4 patients. Because ESI Levels 4 and 5 patients are the least severe cases, they have the lowest priority in treatment bed assignment. Hence, they do not affect the outcomes regarding other patient types when NT method is applied. The number of treatment beds occupied by type 4 patients under NT method is given by

\[ M_{4N}(\lambda, M_N) = \min \left\{ M_N - M_{1N}(\lambda, M_N) - M_{2N}(\lambda, M_N) - M_{3N}(\lambda, M_N), \frac{\kappa_{4N}(\lambda)}{\mu_{4N}} \right\}, \quad (37) \]

where \( \kappa_{4N}(\lambda) = \lambda_4 \). Then, the rates of type 4 patient discharges and abandonments from treatment bed queues can be computed as given in (7) by using \( i = 4 \). Similarly, when PT is applied, the rates of abandonments from PT step queue can be computed by including type 4 patients in the summation on the right hand-side of (8) and computing other outcomes similar to prescribed for type 3 patients in (9)–(14) assuming ESI Levels 4 and 5 patients are all triage patients.

A.2. Selective PT

We next discuss how to make triage method decisions when some of the patients bypass the PT step. This extension also provides ED managers an analytical tool on how to allocate the limited PT step capacity to different patient groups.

Let \( f_i \) be the fraction of type \( i \) patients who are directed to PT step by the triage nurse, \( i = 1, 2, 3 \). While \( f_i \) proportion of type \( i \) patients follow the patient flow as shown in Figure 2, \( (1 - f_i) \) proportion that are not directed to the PT area are placed in treatment bed queue after NT step as shown in Figure 1. In this case, the fluid approximations under PT method can be modified as
follows: When patients are evaluated at PT step on a FCFS basis, the rate at which type $i$ patients abandon the PT step queue is

$$b_{iP}(\lambda) = f_i \gamma_i \left( \lambda - \left( \sum_{i=1}^{3} \frac{f_i \gamma_i}{\delta_i} \right)^{-1} \right)^{+}. \quad (38)$$

By using $b_{iP}$ as given in (38), the rate of type $i$ patient discharges before being placed in a treatment bed queue, $s_{iP}$, and the rate at which type $i$ patients join treatment bed queue, $\kappa_{iP}$, are calculated as given in (9) and (10) by replacing $\lambda_i$ in (9) by $f_i \lambda_i$. We remind that all type 3 patients who go through PT step are discharged without being sent to treatment step queue. Hence, all type 3 patients in treatment bed queue are those who are triaged at NT step only. However, type 1 and 2 patients in the treatment bed queue may or may not have been triaged at PT step. We define $\beta_i$ to be the fraction of type $i$ patients who are triaged at PT step among all type $i$ patients that are sent to treatment bed queues, where $\beta_i$ is

$$\beta_i = 1 - \frac{(1 - f_i) \lambda_i}{\kappa_{iP}(\lambda)}, \quad i = 1, 2. \quad (39)$$

Additionally, we define $\tilde{\mu}_{iP}$, the average treatment rate for type $i$ patients who are placed in the treatment step queue, as

$$\tilde{\mu}_{iP} = \frac{\mu_{iN} \mu_{iP}}{\beta_i \mu_{iN} + (1 - \beta_i) \mu_{iP}} \quad \text{for } i = 1, 2, \quad \tilde{\mu}_{3P} = \mu_{3N}. \quad (40)$$

The number of treatment beds occupied by type 1 and 2 patients, $M_{1P}$ and $M_{2P}$, are computed as given in (11) by replacing $\mu_{iP}$ by $\tilde{\mu}_{iP}$. The number of treatment beds occupied by type 3 patients is

$$M_{3P}(\lambda, M_P) = \min \left\{ M_P - M_{1P}(\lambda, M_P) - M_{2P}(\lambda, M_P), \frac{\kappa_{3P}(\lambda)}{\tilde{\mu}_{3P}} \right\}. \quad (41)$$

Then, the rate at which type $i$ patients are discharged from treatment step, $d_{iP}$, and abandon the treatment step queue, $a_{iP}$, are obtained as given in (13) and (14) by again replacing $\mu_{iP}$ by $\tilde{\mu}_{iP}$. Finally, we modify the revenue and cost parameters to reflect the average values for all type $i$ patients who are placed in treatment bed queues. We define $\tilde{r}_{iP}$ and $\tilde{w}_{iP}$ as

$$\tilde{r}_{iP} = \beta_i r_{iP} + (1 - \beta_i) r_{iN}, \quad \text{for } i = 1, 2, \quad \tilde{r}_{3P} = \frac{s_{3P}(\lambda) r_{1P} + d_{3P}(\lambda) r_{1N}}{s_{3P}(\lambda) + d_{3P}(\lambda)},$$

$$\tilde{w}_{iP} = \beta_i w_{iP} + (1 - \beta_i) w_{iN}, \quad \text{for } i = 1, 2, \quad \tilde{w}_{3P} = w_{3N}. \quad (42)$$

The approximate profit under PT method for all $f_i \in [0, 1], \ i = 1, 2, 3$, can be calculated by replacing $r_{iP}$ and $w_{iP}$ in (16) by $\tilde{r}_{iP}$ and $\tilde{w}_{iP}$. Alternatively, if PT method is implemented by prioritizing patients based on their severity when transferring them to the PT step, the rate of type $i$ patient abandonments from PT step queue, $b_{iP}$, is obtained as prescribed in Remark 2 when $\lambda_i$ in (15) is replaced by $f_i \lambda_i, \ i = 1, 2, 3$. 
Remark 5. We highlight that the model presented for selective PT also captures three well-known implementations of PT method. First, when \( f_i = 1 \) for \( i = 1, 2, 3 \) so that all patients are triaged by a provider as in the “Rapid Medical Evaluation” method, the approximations presented in this section are the same as those provided in §4.1. Second, when \( f_1 = f_2 = 0 \) and \( f_3 = 1 \), only the triage patients are directed to PT area; and hence, the triage provider is dedicated to treating and discharging patients from PT step. This case represents the “See and Treat” model discussed in §3.2. Finally, when \( f_1 = f_2 = 1 \) and \( f_3 = 0 \), only the patients who need further treatment in the treatment step after the PT step are seen by the triage provider, which represents the “PT for Severe Patients” model explained in §3.1.

A.3. Misclassification

Our solution approach presented in §4.1 assumes that patients can be correctly classified according to their types under all triage methods. However, patients’ severity or resource needs may not always be correctly identified at triage. Misclassification of patient type leads to assigning patients to inaccurate priority levels, which can result in undesired cases such as prioritizing non-urgent patients while more urgent ones abandon due to long waits. Because providers are more qualified in assessing patient needs than triage nurses, having patients examined by a provider sooner in the process (as in PT method) can potentially affect misclassifications. We next discuss how to modify our approximations to incorporate the effect of triage method decisions on misclassifications.

We assume that providers can perfectly identify patient type; and hence, under PT method misclassification does not occur beyond PT step. The idea behind this assumption is to take PT method as the base case and approximate the misclassifications under NT relative to PT. In this setting, when patients are directed from NT step to PT step on a FCFS basis as in the setting we analyze in §4.1, misclassification does not affect the outcomes. Hence, even when misclassification is considered, the performance of PT method can be approximated as prescribed in §4.1.

When NT method is applied, the approximations defined in §4.1.1 need to be modified. We define \( \alpha_{ij} \) to be the fraction of type \( i \) patients who are classified as type \( j \) patients by the triage nurse, \( i, j \in \{1, 2, 3\} \). Hence, for example, \( \alpha_{11} \) is the proportion of type 1 patients, \( \alpha_{21} \) is the proportion of type 2 patients and \( \alpha_{31} \) is the proportion of type 3 patients are labeled as type 1 patients and are placed in type 1 patient queue. Then, the fraction of type \( i \) patients in type \( j \) patient queue, \( \Delta_{ij} \), are given by

\[
\Delta_{ij} = \frac{\alpha_{ij} \gamma_i}{\sum_{k=1}^{3} \alpha_{kj} \gamma_k}, \quad i, j = 1, 2, 3.
\]
Because each treatment bed queue potentially contains patients of different types, we define the average treatment rate for patients who are placed in type $j$ patient queue, $\tilde{\mu}_{jN}$, as

$$\tilde{\mu}_{jN} = \left( \frac{\sum_{i=1}^{3} \Delta_{ij}}{\mu_{1N}} \right)^{-1}. \quad (45)$$

Then, the number of treatment beds that are occupied by patients who are labeled as type $j$ is

$$\tilde{M}_{1N}(\lambda, M_N) = \min \left\{ M_N, \frac{\sum_{i=1}^{3} \alpha_{i1} \kappa_{iN}(\lambda)}{\mu_{1N}} \right\}, \quad (46)$$

$$\tilde{M}_{2N}(\lambda, M_N) = \min \left\{ M_N - \tilde{M}_{1N}(\lambda, M_N), \frac{\sum_{i=1}^{3} \alpha_{i2} \kappa_{iN}(\lambda)}{\mu_{2N}} \right\}, \quad (47)$$

$$\tilde{M}_{3N}(\lambda, M_N) = \min \left\{ M_N - \tilde{M}_{1N}(\lambda, M_N) - \tilde{M}_{2N}(\lambda, M_N), \frac{\sum_{i=1}^{3} \alpha_{i3} \kappa_{iN}(\lambda)}{\mu_{3N}} \right\}. \quad (48)$$

Because $\Delta_{ij}$ fraction of the departures from type $j$ patient queue are type $i$ patients, the rate of type $i$ patient departures is given by

$$d_iN(\lambda, M_N) = \sum_{j=1}^{3} \Delta_{ij} \tilde{\mu}_{jN} \tilde{M}_{jN}(\lambda, M_N), \quad (49)$$

and the rate of type $i$ patient abandonments from treatment bed queues, $a_{iN}$, is as given in (7).

**Remark 6.** If patients are directed from NT step to PT step based on their severity as discussed in Remark 2, misclassification hurts the performance of PT method as well. Misclassification under PT method can be incorporated into our model similarly to the case with NT as discussed above. Define

$$\tilde{\delta}_j = \left( \frac{\sum_{i=1}^{3} \Delta_{ij}}{\delta_i} \right)^{-1}. \quad (50)$$

Then, the rate at which patients who are labeled as type $i$ are triaged at PT step, $\tilde{d}_i$, by

$$\tilde{d}_1(\lambda) = \min \left\{ \sum_{k=1}^{3} \alpha_{k1} \lambda_k, \frac{\tilde{\delta}_1}{\tilde{d}_1} \right\}, \quad \tilde{d}_2(\lambda) = \min \left\{ \sum_{k=1}^{3} \alpha_{k2} \lambda_k, \left( 1 - \frac{\sum_{k=1}^{3} \alpha_{k1} \lambda_k}{\delta_1} \right) \frac{\tilde{\delta}_2}{\tilde{d}_2} \right\}, \quad (51)$$

$$\tilde{d}_3(\lambda) = \min \left\{ \sum_{k=1}^{3} \alpha_{k3} \lambda_k, \left( 1 - \frac{\sum_{k=1}^{3} \alpha_{k1} \lambda_k}{\delta_1} - \frac{\sum_{k=1}^{3} \alpha_{k2} \lambda_k}{\delta_2} \right) \frac{\tilde{\delta}_3}{\tilde{d}_3} \right\}. \quad (52)$$

When severity-based prioritization is used in transferring patients to PT step, the abandonment rates from PT step queues are

$$b_{iP}(\lambda) = \lambda_i - \sum_{j=1}^{3} \Delta_{ij} \tilde{d}_i(\lambda), \quad (53)$$

where $\Delta_{ij}$ is as given in (44). Also, the rate of discharges from PT area and transfers to treatment bed queues for type $i$ patients, $s_{iP}$ and $\kappa_{iP}$, are given by (9) and (10). The rates of abandonments and discharges from treatment step are again computed as given in (11)–(14).
A.4. Alternative objective function

As an alternative to the objective of profit maximization, we next consider the minimization of the number of abandonments. PT leads to an increase in treatment step capacity, and hence decreases the abandonments from the treatment bed queue if the staffing level and the number of beds are not reduced when PT is implemented. However, due to limited PT step capacity, abandonments prior to joining the treatment bed queue can potentially offset the decrease in abandonments from the treatment bed queue. Hence, when PT step is the bottleneck, it is not clear whether PT method improves the number of left-without-being-seen (LWBS) patients.

If the ED manager’s goal is to keep the number of LWBS patients to minimum, the objective is to obtain a triage policy $\pi$ that minimizes

$$
\Phi^\pi(T) = \sum_{i=1}^{3} \sum_{j \in \{N,P\}} (\mathbb{E}[B_{ij}^\pi(T)] + \mathbb{E}[E_{ij}^\pi(T)]).
$$

(54)

Depending on the setting considered, the hospital may regard the patients who abandon the ED between PT step and treatment step as abandonments against medical advice (AMA) and may not count them as LWBS patients. In such cases, the terms $\mathbb{E}[E_{ij}^\pi(T)]$ should be excluded from the objective function or be given a different weight in (54) for all $i = 1, 2, 3$.

By using the approximations prescribed in §4.1, we approximate the rate of change in the number of LWBS patients by

$$
\Theta^N(\lambda, M_N) = \sum_{i=1}^{3} a_i N(\lambda, M_N),
$$

$$
\Theta^P(\lambda, M_P) = \sum_{i=1}^{3} (b_i P(\lambda) + a_i P(\lambda, M_P)).
$$

(55)

Our proposed solution is to use triage method that has less abandonment cost.

A.5. Acuity level-based prioritization in treatment bed queue

In some EDs, patients can be prioritized only according to their acuity level in treatment bed queue (i.e., ESI level) and are treated on a FCFS basis within each acuity level. In our model, this translates into type 1 patients having higher (non-preemptive) priority over type 2 and 3 patients, who are of the same priority class. In this case, the performance of PT method can still be approximated as proposed in §4.1.2. Under NT method, the outcomes regarding type 1 patients are obtained as in §4.1.1, but the number of treatment beds occupied by type 2 and 3 patients should be modified as

$$
M_{iN}(\lambda, M_N) = \frac{\gamma_i}{\gamma_2 + \gamma_3} \min \left\{ M_N - M_{iN}(\lambda, M_N), \frac{\kappa_{2N}(\lambda)}{\mu_{2N}} + \frac{\kappa_{3N}(\lambda)}{\mu_{3N}} \right\}, \ i = 2, 3.
$$

(56)

The abandonments and discharges of type 2 and 3 patients can be approximated by (7) by using $M_{iN}(\lambda, M_N)$ as given in (56).
B. Proof of Proposition 1

Proof: We compare the fluid approximations discussed in §4.1 for the profit function of ED X, $\Theta^N$ and $\Theta^P$, in (16) when NT and PT are applied in the ED and obtain the intervals of arrival rate $\lambda$ in which $\Theta^P(\lambda, M) > \Theta^N(\lambda, M)$. Because the capacity at both NT and PT steps are sufficient in ED X, we assume $\kappa_{ij}(\lambda) = \lambda_i$ for $i = 2, 3$ and $j = N, P$. The functions $\Theta^N$ and $\Theta^P$ are piecewise linear, and each one has a change in slope in at most two points. Therefore, we compare them in between these four points.

Interval 1 is defined by $0 \leq \lambda \leq \frac{M_{\mu_2 N P_3 N}}{\gamma \mu_2 N + (1 - \gamma) \mu_3 N}$. In this interval $a_{iN}(\lambda, M) = a_{iP}(\lambda, M) = 0$, $i = 2, 3$, that is, there are no abandonments from the system under either triage method. Interval 2 is defined by $\frac{M_{\mu_2 N P_3 N}}{\gamma \mu_2 N + (1 - \gamma) \mu_3 N} < \lambda < \frac{M_{\mu_2 N}}{(1 - \gamma)}$. In this interval $a_{2N}(\lambda, M) = 0$ and $a_{3N}(\lambda, M) > 0$; hence, the treatment capacity is sufficient for type 2 patients, but some of type 3 patients abandon when NT is applied, and $a_{iP}(\lambda, M) = 0$ for $i = 2, 3$; there are no abandonments when PT is applied. Interval 3 is defined by $\frac{M_{\mu_2 N}}{(1 - \gamma)} < \lambda < \frac{M_{\mu_2 P}}{(1 - \gamma)}$, $a_{2N}(\lambda, M) > 0$ and $a_{3N}(\lambda, M) > 0$, so some of type 2 patients and all of type 3 patients abandon when NT is applied, and $a_{iP}(\lambda, M) = 0$, all patients can be treated when PT is applied. Finally, interval 4 is defined by $\lambda > \frac{M_{\mu_2 P}}{(1 - \gamma)}$, $a_{2P}(\lambda, M) > 0$, so the treatment capacity is not sufficient for type 2 patients and some of them abandon when PT is applied.

The outline of this proof is as follows: In each of the four intervals of $\lambda$, we first find the necessary and sufficient condition for an intersection of $\Theta^N$ and $\Theta^P$. Then, if this condition is met, we find the arrival rate $\lambda$ at which the intersection occurs.

Interval 1: $0 \leq \lambda \leq \frac{M_{\mu_2 N P_3 N}}{\gamma \mu_2 N + (1 - \gamma) \mu_3 N}$.

In interval 1, by (4)–(14), the functions $\Theta^N$ and $\Theta^P$ as defined in (16) become:

$$
\Theta^N(\lambda, M) = r_2 (1 - \gamma) \lambda + r_3 \gamma \lambda - c_N,
\Theta^P(\lambda, M) = r_2 (1 - \gamma) \lambda + r_3 \gamma \lambda - c_P.
$$

Thus, since $c_N < c_P$, we have

$$
\Theta^N(\lambda, M) > \Theta^P(\lambda, M) \forall \lambda \in \left[0, \frac{M_{\mu_2 N \mu_3 N}}{\gamma \mu_2 N + (1 - \gamma) \mu_3 N} \right],
$$

which implies that $\Theta^N$ and $\Theta^P$ do not intersect in interval 1.

Interval 2: $\frac{M_{\mu_2 N P_3 N}}{\gamma \mu_2 N + (1 - \gamma) \mu_3 N} < \lambda < \frac{M_{\mu_2 N}}{(1 - \gamma)}$.

By (4)–(14) and (16), $\Theta^N$ and $\Theta^P$ are defined as follows in interval 2:

$$
\Theta^N(\lambda, M) = r_2 (1 - \gamma) \lambda + r_3 \left( M - \frac{(1 - \gamma) \lambda}{\mu_2 N} \right) \mu_3 N - w_3 N \left( \gamma \lambda - \left( M - \frac{(1 - \gamma) \lambda}{\mu_2 N} \right) \mu_3 N \right) - c_N,
$$

(59)
\[ \Theta^p(\lambda, M) = r_2(1 - \gamma)\lambda + r_3\gamma\lambda - c_P. \] 

(60)

By (58), we have

\[ \Theta^N \left( \frac{M\mu_{2N}\mu_{3N}}{\gamma\mu_{2N} + (1 - \gamma)\mu_{3N}}, M \right) > \Theta^P \left( \frac{M\mu_{2N}\mu_{3N}}{\gamma\mu_{2N} + (1 - \gamma)\mu_{3N}}, M \right). \] 

(61)

Because \( \Theta^N \) and \( \Theta^P \) are linear in interval 2 and by (59)–(61), a necessary and sufficient condition for an intersection in interval 2 is

\[ \Theta^N \left( \frac{M\mu_{2N}}{(1 - \gamma)}, M \right) = r_2M\mu_{2N} - w_{3N}\gamma\mu_{2N} - c_N \]
\[ \leq \Theta^P \left( \frac{M\mu_{2N}}{(1 - \gamma)}, M \right) = r_2M\mu_{2N} + r_3\gamma\mu_{2N} - c_P, \]

(62)

which is equivalent to Condition 1 in (19). Thus, if and only if (19) holds, there exists a switching arrival rate \( \lambda_{s1} \), i.e., \( \Theta^N(\lambda_{s1}, M) = \Theta^P(\lambda_{s1}, M) \), that is defined as follows by (59) and (60):

\[ \lambda_{s1} = \frac{\mu_{2N}(M\mu_{3N}(r_3 + w_{3N}) + c_P - c_N)}{(r_3 + w_{3N})(\gamma\mu_{2N} + (1 - \gamma)\mu_{3N})}. \]

(63)

This implies that if (19) holds and \( \lambda_{s1} < \frac{M\mu_{2N}}{(1 - \gamma)} \), we have

\[ \Theta^N(\lambda, M) \geq \Theta^P(\lambda, M) \text{ for } \lambda \in \left[ \frac{M\mu_{2N}\mu_{3N}}{\gamma\mu_{2N} + (1 - \gamma)\mu_{3N}}, \lambda_{s1} \right], \]
\[ \Theta^N(\lambda, M) < \Theta^P(\lambda, M) \text{ for } \lambda \in \left( \lambda_{s1}, \frac{M\mu_{2N}}{(1 - \gamma)} \right). \]

If (19) holds and \( \lambda_{s1} = \frac{M\mu_{2N}}{(1 - \gamma)} \),

\[ \Theta^N(\lambda, M) > \Theta^P(\lambda, M) \text{ for } \lambda \in \left[ \frac{M\mu_{2N}\mu_{3N}}{\gamma\mu_{2N} + (1 - \gamma)\mu_{3N}}, \frac{M\mu_{2N}}{(1 - \gamma)} \right], \text{ and } \Theta^N \left( \frac{M\mu_{2N}}{(1 - \gamma)}, M \right) = \Theta^P \left( \frac{M\mu_{2N}}{(1 - \gamma)}, M \right). \]

In the other case in which (19) is not satisfied, by (61), \( \Theta^N \) and \( \Theta^P \) being linear in interval 2 and that (62) fails to hold, there does not exist a switching arrival rate, and \( \Theta^N(\lambda, M) > \Theta^P(\lambda, M) \) for \( \lambda \leq \frac{M\mu_{2N}}{(1 - \gamma)}. \)

**Interval 3:** \( \frac{M\mu_{2N}}{(1 - \gamma)} < \lambda \leq \frac{M\mu_{2P}}{(1 - \gamma)} \)

By (4)–(14) and (16), \( \Theta^N \) and \( \Theta^P \) and their first derivatives are defined as follows in interval 3:

\[ \Theta^N(\lambda, M) = r_2M\mu_{2N} - w_{2N}(1 - \gamma)\lambda - M\mu_{2N}) - w_{3N}\gamma\lambda - c_N, \]

(64)

\[ \Theta^P(\lambda, M) = r_2(1 - \gamma)\lambda + r_3\gamma\lambda - c_P, \]

(65)

\[ \frac{\partial \Theta^N(\lambda, M)}{\partial \lambda} = -(1 - \gamma)w_{2N} - \gamma w_{3N}, \]
\[ \frac{\partial \Theta^P(\lambda, M)}{\partial \lambda} = (1 - \gamma)r_2 + \gamma r_3. \]
First, by $\gamma \in [0, 1]$, $r_2, r_3 > 0$ and $w_{2N}, w_{3N} \geq 0$, we have
\[
\frac{\partial \Theta^p(\lambda, M)}{\partial \lambda} > \frac{\partial \Theta^N(\lambda, M)}{\partial \lambda}.
\] (66)
If (19) is satisfied, by (62) and (66), $\Theta^p > \Theta^N$ in interval 3. If (19) does not hold, then $\Theta^p$ and $\Theta^N$ may or may not intersect. Because (62) fails to hold in this case, by (64), (65) and the linearity of $\Theta^N$ and $\Theta^p$ in interval 3, an intersection will occur if and only if the following holds:
\[
\Theta^N\left(\frac{M\mu_{2P}}{(1-\gamma)}, M\right) = r_2M\mu_{2N} - w_{2N}M(\mu_{2P} - \mu_{2N}) - w_{3N}\gamma M\mu_{2P} + c_N
\]
\[
\leq \Theta^p\left(\frac{M\mu_{2P}}{(1-\gamma)}, M\right) = r_2M\mu_{2P} + r_3\gamma M\mu_{2P} - c_p,
\]
which is equivalent to Condition 2 in (20). Then, if (19) fails and (20) is satisfied, there exists a unique switching arrival rate $\lambda_{s2}$, i.e., $\Theta^N(\lambda_{s2}, M) = \Theta^p(\lambda_{s2}, M)$, in interval 3, which is found by using (64) and (65) as
\[
\lambda_{s2} = \frac{(r_2 + w_{2N})M\mu_{2N} + c_p - c_N}{\gamma(r_3 + w_{3N}) + (1-\gamma)(r_2 + w_{2N})}.
\] (67)
Thus, if (19) fails, (20) holds and $\lambda_{s2} < \frac{M\mu_{2P}}{(1-\gamma)}$, we have $\Theta^N(\lambda, M) \geq \Theta^p(\lambda, M)$ for $\lambda \in \left[\frac{M\mu_{2N}}{(1-\gamma)}, \lambda_{s2}\right]$ and $\Theta^N(\lambda, M) < \Theta^p(\lambda, M)$ for $\lambda \in \left(\lambda_{s2}, \frac{M\mu_{2P}}{(1-\gamma)}\right]$. If (19) fails, (20) holds and $\lambda_{s2} = \frac{M\mu_{2P}}{(1-\gamma)}$, $\Theta^N(\lambda, M) > \Theta^p(\lambda, M)$ for $\lambda \in \left[\frac{M\mu_{2N}}{(1-\gamma)}, \frac{M\mu_{2P}}{(1-\gamma)}\right]$ and $\Theta^N(\frac{M\mu_{2P}}{(1-\gamma)}, M) = \Theta^p\left(\frac{M\mu_{2P}}{(1-\gamma)}, M\right)$. Finally, if (19) and (20) are not satisfied, because (62) will also not be satisfied, $\Theta^N$ and $\Theta^p$ do not intersect and $\Theta^N(\lambda, M) > \Theta^p(\lambda, M)$ for $\lambda \leq \frac{M\mu_{2P}}{(1-\gamma)}$. Also, the cases analyzed for interval 3 imply that unless both (19) and (20) fail, the following holds:
\[
\Theta^N\left(\frac{M\mu_{2P}}{(1-\gamma)}, M\right) \leq \Theta^p\left(\frac{M\mu_{2P}}{(1-\gamma)}, M\right),
\] (68)
which will be used later in this proof.

Interval 4: $\lambda > \frac{M\mu_{2P}}{(1-\gamma)}$.

By (4)–(16), $\Theta^N$ and $\Theta^p$ and their first derivatives are defined as follows in interval 4:
\[
\Theta^N(\lambda, M) = r_2M\mu_{2N} - w_{2N}((1-\gamma)\lambda - M\mu_{2N}) - w_{3N}\gamma \lambda - c_N,
\] (69)
\[
\Theta^p(\lambda, M) = r_2M\mu_{2P} + r_3\gamma \lambda - w_{2P}((1-\gamma)\lambda - M\mu_{2P}) - c_p,
\] (70)
\[
\frac{\partial \Theta^N(\lambda, M)}{\partial \lambda} = -(1-\gamma)w_{2N} - \gamma w_{3N},
\] (71)
\[
\frac{\partial \Theta^p(\lambda, M)}{\partial \lambda} = \gamma r_3 - (1-\gamma)w_{2P}.
\]
We first analyze the case for which (19) and (20), and hence (68), both do not hold. In this case, by the fact that (68) fails to hold and the linearity of $\Theta^N$ and $\Theta^p$ in interval 4, a necessary and sufficient condition for an intersection in interval 4 is
\[
\frac{\partial \Theta^p(\lambda, M)}{\partial \lambda} > \frac{\partial \Theta^N(\lambda, M)}{\partial \lambda},
\] (72)
which is equivalent to Condition 3 defined in (21) with strict inequality. If (72) is satisfied, there exists a unique switching arrival rate $\lambda_{s3}$ in interval 4 such that $\Theta^N(\lambda_{s3}, M) = \Theta^P(\lambda_{s3}, M)$, which is, by (69) and (70), defined as

$$\lambda_{s3} = \frac{(r_2 + w_{2N})M\mu_{2N} - (r_2 + w_{2P})M\mu_{2P} + c_P - c_N}{\gamma(r_3 + w_{3N}) - (1 - \gamma)(w_{2P} - w_{2N})}. \quad (73)$$

Then, if all of (19), (20) and (72) fail to hold, then $\Theta^N(\lambda, M) \geq \Theta^P(\lambda, M)$, $\forall \lambda \geq 0$. Because (72) is the strict inequality equivalent of (21), this proves part 1 of Proposition 1. If (19) and (20) fail and (72) holds, implying (21) as a strict inequality, $\Theta^P(\lambda, M) > \Theta^N(\lambda, M)$ for $\lambda > \lambda_{s3}$ and $\Theta^P(\lambda, M) \leq \Theta^N(\lambda, M)$ for $\lambda \leq \lambda_{s3}$, which proves part 2(c) of the proposition.

We next consider the cases in which at least one of (19) and (20), and thus (68), holds. Similar to the idea above, for $\Theta^N$ and $\Theta^P$ to intersect in interval 4, the inequality in (72) must be reversed. Thus, (21) failing to hold becomes a necessary and sufficient condition for an intersection, which will occur at $\lambda_{s3}$.

If (19) holds and (21) does not hold, $\Theta^P(\lambda, M) > \Theta^N(\lambda, M)$ for $\lambda \in (\lambda_{s1}, \lambda_{s3})$ and $\Theta^P(\lambda, M) \leq \Theta^N(\lambda, M)$ when $\lambda \notin (\lambda_{s1}, \lambda_{s3})$, which proves part 3(a) of the proposition. Similarly, if (20) holds, and (19) and (21) do not hold, $\Theta^P(\lambda, M) > \Theta^N(\lambda, M)$ for $\lambda \in (\lambda_{s2}, \lambda_{s3})$ and $\Theta^P(\lambda, M) \leq \Theta^N(\lambda, M)$ for $\lambda \notin (\lambda_{s2}, \lambda_{s3})$, which proves part 3(b) of the proposition.

If either one of (19) and (20) holds, because (21) failing to hold is a necessary condition for an intersection in interval 4, when (21) holds, $\Theta^N$ and $\Theta^P$ do not intersect in interval 4, i.e., $\Theta^P > \Theta^N$ in interval 4. This implies by our previous analysis that if (19) and (21) hold, $\Theta^P(\lambda, M) > \Theta^N(\lambda, M)$ for $\lambda > \lambda_{s1}$ and $\Theta^P(\lambda, M) \leq \Theta^N(\lambda, M)$ for $\lambda \leq \lambda_{s1}$, which proves part 2(a) of the proposition. If (19) fails to hold, and (20) and (21) hold, $\Theta^P(\lambda, M) > \Theta^N(\lambda, M)$ for $\lambda > \lambda_{s2}$ and $\Theta^P(\lambda, M) \leq \Theta^N(\lambda, M)$ for $\lambda \leq \lambda_{s2}$, which proves part 2(b) of the proposition. $\square$

C. Impact of alternative implementations in ED X

In this section, we discuss the impact of different implementations of PT (presented in §4.3) on the structure of the optimal policy for ED X. Our main goal is to assess the robustness of our findings for ED X.

i) An alternative way EDs implement PT is to treat ESI Levels 4 and 5 patients at the triage step, and effectively replacing “Fast Track” area with PT. In this case, ESI Levels 4 and 5 patients are treated by a physician at triage, and they wait in a separate area while their tests are being conducted. This waiting area is staffed with nurses to facilitate the treatment process, in a way similar to how Fast Track operates. We compare this implementation of PT to that of NT when
ESI Levels 4 and 5 patients are treated in the Fast Track area, which we assume has sufficient capacity for all ESI Levels 4 and 5 patients.

The structure of the policy remains the same but the cost parameters need to be updated (assuming revenues for ESI Levels 4 and 5 patients are the same under both triage methods). Specifically, the cost of operating the Fast Track should be added to the total cost under NT, and additional staffing costs must be accounted for in the staffing cost under PT. The optimal policy can then be identified as in Proposition 1 if PT has ample capacity and as in Appendix F if it has limited capacity. The abandonment costs for ESI Levels 4 and 5 patients must be accounted for in the latter case based on the priority rule used in PT.

ii) If ED X uses “See and Treat” model, and hence only triage patients are seen at PT step, the proposed solution can still be obtained by Proposition 1 using \( \mu_2^N = \mu_2^P \) and \( w_2^N = w_2^P \). In this case, the proposed triage method changes at exactly one threshold arrival rate. As for the “PT for Severe Patients” implementation, where only treatment patients are seen at PT step, one needs to account for the triage patients being served in the treatment area under PT. In this case, the profit function under PT has a similar shape to that under NT. Assuming \( \mu_2^P \geq \mu_2^N \) and \( w_3^N = w_3^P \) (since triage patients follow the same treatment path under both triage methods in this case), one can show that the structure of the policy remains the same as in Proposition 1 if PT has ample capacity, and as in Appendix F with limited capacity.

iii and v) If the patients might be misclassified under NT (see Appendix A.3 for details) or if the patients are served in a FCFS basis in the treatment stage under NT, the structure of the optimal policy remains the same if the abandonment cost for treatment patients is high enough. Specifically, if \( w_2^N > \frac{\mu_3^N w_3^N}{\mu_2^N} \), it can be shown that the profit under NT is lower in both cases than the case with no misclassifications for \( \lambda > \lambda_0 \).

iv) If the objective of the ED is to minimize the total number of abandonments, we can use the results in Appendix A.4 to determine the optimal triage decision. Figure 5 presents throughputs under NT (\( \Theta^N \)) and PT (\( \Theta^P \)) when the PT step has unlimited capacity. (We use the same notation \( \lambda_0, \lambda_1 \) and \( \lambda_2 \) introduced in Figure 3.) For arrival rates less than \( \lambda_0 \), both triage methods have the same throughput. (Note that the abandonment rate is given by the arrival rate minus the throughput; hence, it is enough to maximize throughput to minimize abandonment rate for a given arrival rate.) Clearly PT has higher throughput if the arrival rate is higher than \( \lambda_0 \).

In the limited capacity case, it can be shown that there is at most one threshold for the arrival rate when the optimal triage method changes as we explain next. If PT has limited capacity, denoted by \( \tilde{\delta} \), then the throughput under PT is the minimum of \( \Theta^P(\lambda) \) and \( \tilde{\delta} \) for a given arrival
rate \( \lambda \geq 0 \). Hence, if \( \bar{\delta} \geq \delta^*_1 \), where \( \delta^*_1 = \Theta^N(\lambda_1) \) the throughput under NT when arrival rate is \( \lambda_1 \), then throughput is higher under PT than that under NT for all arrival rates above \( \lambda_0 \). Similarly, if \( \bar{\delta} \leq \delta^*_0 \), where \( \delta^*_0 = \Theta^N(\lambda_0) \), throughput is lower under PT than that under NT for all arrival rates above \( \bar{\delta} \). If \( \delta^*_0 \leq \bar{\delta} \leq \delta^*_1 \), then the throughput under PT is higher for arrival rates lower than a threshold but higher for NT above that threshold. The value of this threshold, \( \Lambda_1 \), can be explicitly computed from \( \Theta^N(\Lambda_1) = \bar{\delta} \).

Figure 5 Throughput under NT and PT.

D. Proposed policy for ED X when \( \mu_{2P} \leq \mu_{2N} \) and \( c_P > c_N \).

In this case, PT leads to discharging triage patients without using treatment step capacity, but it slows down the treatment for treatment patients. Therefore, the total capacity in the ED may or may not go down when PT is applied. When \( \mu_{2P} \) is smaller, the treatment capacity under PT decreases; and thus, treatment patients start abandoning at smaller arrival rates. For example, Figure 6 illustrates how the solution can change when \( \mu_{2P} \leq \mu_{2N} \) in the case of two intersections as in Figure 3c. When \( \mu_{2P} \) is smaller, \( \Theta^P \) shifts down after treatment patients start abandoning, which leads the second intersection of \( \Theta^P \) and \( \Theta^N \) to occur at a smaller arrival rate \( \Lambda'_2 \) instead of \( \Lambda_2 \) as in Figure 3c. In general, the structure of the solution when \( \mu_{2P} \leq \mu_{2N} \) is similar to the one presented in Proposition 1 in the sense that in most cases \( \Theta^N \) and \( \Theta^P \) can intersect up to two times. However, different from Proposition 1, a special case in which there exists three intersections is also possible. In order to characterize the exact structure of the solution, we use Conditions 4 and 5, which are defined as

\[
\text{Condition 4: } (r_3 + w_{3N}) \frac{M \gamma \mu_{2P}}{(1 - \gamma)} - (r_3 + w_{3N}) M \left( \mu_{3N} - \frac{\mu_{2P} \mu_{3N}}{\mu_{2N}} \right) \geq c_P - c_N, \tag{74}\]

\[
\text{Condition 5: } (r_3 + w_{3N}) \frac{\gamma M \mu_{2N}}{(1 - \gamma)} - (r_2 + w_{2P}) M (\mu_{2N} - \mu_{2P}) \leq c_P - c_N. \tag{75}\]

Condition 4 implies that \( \Theta^P \) is greater than or equal to \( \Theta^N \) when arrival rate is \( \frac{M \mu_{2P}}{(1 - \gamma)} \), at which all patients are treated under PT, but part of triage patients abandon when NT is applied. Condition
5 implies that $\Theta^N$ is greater than or equal to $\Theta^P$ when the arrival rate is $\frac{M\mu_2N}{(1-\gamma)}$, the rate at which all triage patients abandon and all treatment patients are treated when NT is used, and part of treatment patients abandon under PT if $\mu_2P \leq \mu_2N$. Thus, when Condition 5 holds (assuming $\mu_2P \leq \mu_2N$), the total revenue difference between PT and NT is less than the additional staffing cost.

Proposition 2 presents the proposed solution when $\mu_2P \leq \mu_2N$ based on Conditions 3, 4 and 5.

**Proposition 2.** Assume that $\mu_2P \leq \mu_2N$. The optimal solution $j^\ast$ defined in (18) is given by the following.

1. If $\frac{M\mu_2P\mu_2N}{\gamma\mu_2N+(1-\gamma)\mu_2P} < \frac{M\mu_2P}{(1-\gamma)}$,
   (a) $j^\ast(\lambda, M) = N$ for all $\lambda \geq 0$ if Condition 5 holds, and Conditions 3 and 4 fail to hold.
   (b) $j^\ast(\lambda, M) = N$ for $\lambda \leq \Lambda_1$ and $j^\ast(\lambda, M) = P$ for $\lambda > \Lambda_1$, where
      i. $\Lambda_1$ is given in (22) if Conditions 3 and 4 hold, and Condition 5 fails to hold.
      ii. $\Lambda_1 = \frac{\mu_2N((r_2+w_2P)M\mu_2P-(r_3+w_3N)M\mu_3N+c_N-c_P)}{\mu_2N(r_2+w_2P)(1-\gamma)-(r_3+w_3N)(1-\gamma)\mu_3N-(w_3N+r_3)\gamma\mu_2N}$. (76)
         if Condition 3 holds, Condition 4 fails to hold, and Condition 5 either does not hold or holds as an equality.
      iii. $\Lambda_1$ is given in (24) if Conditions 3 and 5 hold as a strict inequality, and Condition 4 fails to hold.
   (c) $j^\ast(\lambda, M) = N$ for $\lambda \leq \Lambda_1$ and $\lambda \geq \Lambda_2$ and $j^\ast(\lambda, M) = P$ for $\Lambda_1 < \lambda < \Lambda_2$, where
      i. $\Lambda_1$ and $\Lambda_2$ are as given in (22) and (24) if Condition 4 holds, and Conditions 3 and 5 do not hold.
ii. $\Lambda_1$ and $\Lambda_2$ are as given in (22) and (76) if Conditions 4 and 5 hold, and Condition 3 does not hold or holds as an equality.

iii. $\Lambda_1$ and $\Lambda_2$ are as given in (76) and (24) if Conditions 3, 4 and 5 fail to hold.

(d) $j^*(\lambda, M) = N$ for $\lambda \leq \Lambda_1$ and for $\Lambda_2 \leq \lambda \leq \Lambda_3$, and $j^*(\lambda, M) = P$ for $\Lambda_1 < \lambda < \Lambda_2$ and $\lambda > \Lambda_3$, where $\Lambda_1$, $\Lambda_2$ and $\Lambda_3$ are given in (22), (76) and (24), if Conditions 4 and 5 hold, and Condition 3 holds as a strict inequality.

2. If $\frac{M_{\mu N_p}}{\gamma_{\mu N} + (1-\gamma)\mu_3} \geq \frac{M_{\mu P}}{(1-\gamma)}$,

(a) $j^*(\lambda, M) = N$ for all $\lambda \geq 0$ if Condition 5 holds and Condition 3 either fails to hold or holds as an equality.

(b) $j^*(\lambda, M) = N$ for $\lambda \leq \Lambda_1$ and $j^*(\lambda, M) = P$ for $\lambda > \Lambda_1$, where

i. $\Lambda_1$ is as given in (76) if Condition 3 holds and Condition 5 fails to hold.

ii. $\Lambda_1$ is as given in (24) if Condition 3 holds as a strict inequality and Condition 5 holds.

(c) $j^*(\lambda, M) = N$ for $\lambda \leq \Lambda_1$ and $\lambda \geq \Lambda_2$, and $j^*(\lambda, M) = P$ for $\Lambda_1 < \lambda < \Lambda_2$ if Conditions 3 and 5 fail to hold, where $\Lambda_1$ and $\Lambda_2$ are given by (76) and (24).

Proof of Proposition 2: When $\mu_2 \leq \mu_2 N$, $\frac{M_{\mu N_p}}{\gamma_{\mu N} + (1-\gamma)\mu_3} < \frac{M_{\mu P}}{(1-\gamma)}$ may or may not hold. We analyze these two cases separately.

When $\frac{M_{\mu N_p}}{\gamma_{\mu N} + (1-\gamma)\mu_3} < \frac{M_{\mu P}}{(1-\gamma)}$ holds

We compare $\Theta^N$ and $\Theta^P$ in four intervals of $\lambda$. Interval 1 is defined by $0 \leq \lambda \leq \frac{M_{\mu N_p}}{\gamma_{\mu N} + (1-\gamma)\mu_3}$, where $a_{iN}(\lambda, M) = a_{iP}(\lambda, M) = 0$, $i = 2, 3$. Interval 2 is defined by $\frac{M_{\mu N_p}}{\gamma_{\mu N} + (1-\gamma)\mu_3} < \lambda \leq \frac{M_{\mu P}}{(1-\gamma)}$, in which $a_{2j}(\lambda, M) = 0$, $j = N, P$, and $a_{3N}(\lambda, M) > 0$. Interval 3 is defined by $\frac{M_{\mu P}}{(1-\gamma)} < \lambda \leq \frac{M_{\mu N}}{(1-\gamma)}$. In this interval, $a_{2P}(\lambda, M) > 0$, $a_{2N}(\lambda, M) = 0$ and $a_{3N}(\lambda, M) > 0$. Interval 4 is defined by $\lambda > \frac{M_{\mu N}}{(1-\gamma)}$, where $a_{2P}(\lambda, M) > 0$, $a_{2N}(\lambda, M) > 0$ and $a_{3N}(\lambda, M) > 0$.

By (4)–(14), (16) and that $c_P > c_N$, it is trivial to show that

$$\Theta^N(\lambda, M) > \Theta^P(\lambda, M) \text{ for } \lambda \in \left[0, \frac{M_{\mu N_p}}{\gamma_{\mu N} + (1-\gamma)\mu_3}\right].$$

(77)

We next analyze potential intersections in intervals 2, 3 and 4.

Interval 2: $\frac{M_{\mu N_p}}{\gamma_{\mu N} + (1-\gamma)\mu_3} < \lambda \leq \frac{M_{\mu P}}{(1-\gamma)}$.

By (4)–(14) and (16), $\Theta^N$ and $\Theta^P$ are defined as in (59) and (60). Then, by the linearity of $\Theta^N$ and $\Theta^P$ in interval 3, the necessary and sufficient condition for an intersection in interval 2 is:

$$\Theta^N\left(\frac{M_{\mu P}}{(1-\gamma)}, M\right) = r_2 M_{\mu 2P} + r_3 \left(M - \frac{M_{\mu 2P}}{\mu_2 N}\right) \mu_3 N - w_3 N \left(\gamma M_{\mu 2P}\frac{\gamma_{\mu N_p}}{\gamma_{\mu N} + (1-\gamma)\mu_3}\right) - c_N$$

$$\leq \Theta^P\left(\frac{M_{\mu P}}{(1-\gamma)}, M\right) = r_2 M_{\mu 2P} + r_3 \gamma \frac{M_{\mu 2P}}{(1-\gamma)} - c_P,$$  

(78)
which is equivalent to Condition 4 in (74). Thus, if and only if Condition 4 holds, there exists a switching arrival rate \( \lambda_{s4} \), i.e., \( \Theta^N(\lambda_{s4}, M) = \Theta^P(\lambda_{s4}, M) \), that is given in (63).

**Interval 3:** \( \frac{\mu_{2P}}{(1-\gamma)} \leq \lambda \leq \frac{\mu_{2N}}{(1-\gamma)} \)

By (4)–(14) and (16), \( \Theta^N \) and \( \Theta^P \) are defined as in (59) and (70).

If Condition 4 holds so that \( \Theta^P \left( \frac{\mu_{2P}}{(1-\gamma)}, M \right) \geq \Theta^N \left( \frac{\mu_{2P}}{(1-\gamma)}, M \right) \), there exists an intersection in interval 3 if and only if the following holds:

\[
\Theta^N \left( \frac{\mu_{2N}}{(1-\gamma)}, M \right) = r_2M\mu_{2N} - w_3\gamma \frac{\mu_{2N}}{(1-\gamma)} - c_N \\
\geq \Theta^P \left( \frac{\mu_{2N}}{(1-\gamma)}, M \right) = r_2M\mu_{2P} + r_3\gamma \frac{\mu_{2N}}{(1-\gamma)} - w_2(M\mu_{2N} - M\mu_{2P}) - c_P,
\]

which is equivalent to Condition 5 in (75). Similarly, if Condition 4 does not hold so that \( \Theta^P \left( \frac{\mu_{2P}}{(1-\gamma)}, M \right) < \Theta^N \left( \frac{\mu_{2P}}{(1-\gamma)}, M \right) \), Condition 5 failing to hold becomes a necessary and sufficient condition for an intersection in interval 3.

The switching arrival rate (if it exists) in interval 3, \( \lambda_{s5} \), i.e., \( \Theta^N(\lambda_{s5}, M) = \Theta^P(\lambda_{s5}, M) \), is defined as follows by (59) and (70).

\[
\lambda_{s5} = \frac{\mu_{2N}(r_2 + w_2\lambda_{s5})M\mu_{2P} - (r_3 + w_3\lambda_{s5})M\mu_{3N} + c_N - c_P}{\mu_{2N}(r_2 + w_2\lambda_{s5})(1-\gamma) - (r_3 + w_3\lambda_{s5})(1-\gamma)\mu_{3N} - (w_3\lambda_{s5} + r_3)\gamma\mu_{2N}}.
\] (80)

**Interval 4:** \( \lambda > \frac{\mu_{2N}}{(1-\gamma)} \).

By (4)–(14) and (16), \( \Theta^N \) and \( \Theta^P \) are defined as in (69) and (70). If Condition 5 holds so that \( \Theta^N \left( \frac{\mu_{2N}}{(1-\gamma)}, M \right) \geq \Theta^P \left( \frac{\mu_{2N}}{(1-\gamma)}, M \right) \), the necessary and sufficient condition for an intersection in interval 4 is \( \frac{\partial\Theta^N(\lambda, M)}{\partial \lambda} > \frac{\partial\Theta^P(\lambda, M)}{\partial \lambda} \), which is equivalent to Condition 3 defined in (21) with strict inequality.

Similarly, if Condition 5 does not hold so that \( \Theta^N \left( \frac{\mu_{2N}}{(1-\gamma)}, M \right) < \Theta^P \left( \frac{\mu_{2N}}{(1-\gamma)}, M \right) \), Condition 3 failing to hold becomes a necessary and sufficient condition for an intersection in interval 4. In these cases, the switching arrival rate \( \lambda_{s6} \), i.e., \( \Theta^N(\lambda_{s6}, M) = \Theta^P(\lambda_{s6}, M) \), is given by (73).

If only Condition 5 holds, \( \Theta^N(\lambda, M) \geq \Theta^P(\lambda, M) \) for all \( \lambda \geq 0 \). If only Conditions 3 and 4 hold, \( \Theta^N(\lambda, M) \geq \Theta^P(\lambda, M) \) for \( \lambda \in [0, \lambda_{s4}] \) and \( \Theta^P(\lambda, M) > \Theta^N(\lambda, M) \) for \( \lambda \in (\lambda_{s4}, \infty) \). If only Conditions 3 and 5 hold, \( \Theta^N(\lambda, M) \geq \Theta^P(\lambda, M) \) for \( \lambda \in [0, \lambda_{s6}] \) and \( \Theta^P(\lambda, M) > \Theta^N(\lambda, M) \) for \( \lambda \in (\lambda_{s6}, \infty) \).

If only Condition 3 holds, \( \Theta^N(\lambda, M) \geq \Theta^P(\lambda, M) \) for \( \lambda \in [0, \lambda_{s5}] \) and \( \Theta^P(\lambda, M) > \Theta^N(\lambda, M) \) for \( \lambda \in (\lambda_{s5}, \infty) \).

If only Condition 4 holds, \( \Theta^N(\lambda, M) \geq \Theta^P(\lambda, M) \) for \( \lambda \in [0, \lambda_{s4}] \) and \( \lambda \in [\lambda_{s6}, \infty) \), and \( \Theta^P(\lambda, M) > \Theta^N(\lambda, M) \) for \( \lambda \in (\lambda_{s4}, \lambda_{s6}) \). If only Conditions 4 and 5 hold, \( \Theta^N(\lambda, M) \geq \Theta^P(\lambda, M) \) for \( \lambda \in [0, \lambda_{s4}] \) and \( \lambda \in [\lambda_{s5}, \infty) \), and \( \Theta^P(\lambda, M) > \Theta^N(\lambda, M) \) for \( \lambda \in (\lambda_{s4}, \lambda_{s5}) \). If none of Conditions 3, 4 and 5 holds, \( \Theta^N(\lambda, M) \geq \Theta^P(\lambda, M) \) for \( \lambda \in [0, \lambda_{s5}] \) and \( \lambda \in [\lambda_{s6}, \infty) \), and \( \Theta^P(\lambda, M) > \Theta^N(\lambda, M) \) for
\( \lambda \in (\lambda_{s5}, \lambda_{s6}) \). When all of Conditions 3, 4 and 5 hold, \( \Theta^N(\lambda, M) \geq \Theta^P(\lambda, M) \) for \( \lambda \in [0, \lambda_{s4}] \) and 
\( \lambda \in [\lambda_{s5}, \lambda_{s6}] \), and \( \Theta^P(\lambda, M) > \Theta^N(\lambda, M) \) for \( \lambda \in (\lambda_{s4}, \lambda_{s5}) \) and \( \lambda \in (\lambda_{s6}, \infty) \).

**When \( \frac{M_{\mu2N}M_{\mu3N}}{\gamma_{\mu2N}+(1-\gamma)\mu_{3N}} \geq \frac{M_{\mu2N}}{(1-\gamma)} \) holds**

With the same reasoning as in the previous proofs, we compare \( \Theta^N \) and \( \Theta^P \) in four intervals of \( \lambda \): 
- \( \left(0, \frac{M_{\mu2P}}{(1-\gamma)}\right) \), 
- \( \left(\frac{M_{\mu2P}}{(1-\gamma)}, \frac{M_{\mu2P}M_{\mu3N}}{\gamma_{\mu2N}+(1-\gamma)\mu_{3N}}\right) \), 
- \( \left(\frac{M_{\mu2P}M_{\mu3N}}{\gamma_{\mu2N}+(1-\gamma)\mu_{3N}}, \frac{M_{\mu2N}}{(1-\gamma)}\right) \), 
- \( (\frac{M_{\mu2N}}{(1-\gamma)}, \infty) \). It is trivial to show that by By (4)–(14), (16) and that the switching arrival rate (if it exists) in interval 3 is defined by (73).

When \( \frac{M_{\mu2N}M_{\mu3N}}{\gamma_{\mu2N}+(1-\gamma)\mu_{3N}} \geq \frac{M_{\mu2N}}{(1-\gamma)} \), the necessary and sufficient condition for an intersection in interval 4 is 
\( \Theta^N(\lambda, M) > \Theta^P(\lambda, M) \) for \( \lambda \in \left(0, \frac{M_{\mu2P}}{(1-\gamma)}\right) \) .

Also, in interval 2 in which 
\( \left(\frac{M_{\mu2P}}{(1-\gamma)}, \frac{M_{\mu2P}M_{\mu3N}}{\gamma_{\mu2N}+(1-\gamma)\mu_{3N}}\right) \), \( \Theta^N \) and \( \Theta^P \) are given by (57) and (70). Then, in interval 2, we have
\[
\frac{\partial \Theta^N(\lambda, M)}{\partial \lambda} = r_2(1-\gamma) + r_3\gamma > \frac{\partial \Theta^P(\lambda, M)}{\partial \lambda} = r_3\gamma - w_{2P}(1-\gamma).
\]

By (81) and (82), \( \Theta^N \) and \( \Theta^P \) do not intersect in intervals 1 and 2 and we have
\[
\Theta^N\left(\frac{M_{\mu2P}M_{\mu3N}}{\gamma_{\mu2N}+(1-\gamma)\mu_{3N}}, M\right) > \Theta^P\left(\frac{M_{\mu2P}M_{\mu3N}}{\gamma_{\mu2N}+(1-\gamma)\mu_{3N}}, M\right).
\]

We next analyze intervals 3 and 4.

**Interval 3:** 
\( \frac{M_{\mu2P}M_{\mu3N}}{\gamma_{\mu2N}+(1-\gamma)\mu_{3N}} < \lambda \leq \frac{M_{\mu2N}}{(1-\gamma)} \).

In this interval \( \Theta^N \) and \( \Theta^P \) are given by (59) and (70). Then, for an intersection in interval 3, the necessary and sufficient condition is 
\( \Theta^N\left(\frac{M_{\mu2N}}{(1-\gamma)}, M\right) \leq \Theta^P\left(\frac{M_{\mu2N}}{(1-\gamma)}, M\right) \), which is Condition 5 in (75) with reverse direction. The switching arrival rate (if it exists) in interval 3, \( \lambda_{s7} \), i.e., \( \Theta^N(\lambda_{s7}, M) = \Theta^P(\lambda_{s7}, M) \), is given by (80).

**Interval 4:** 
\( \lambda > \frac{M_{\mu2N}}{(1-\gamma)} \).

If there does not exist an intersection in interval 3 so that \( \Theta^N\left(\frac{M_{\mu2N}}{(1-\gamma)}, M\right) \geq \Theta^P\left(\frac{M_{\mu2N}}{(1-\gamma)}, M\right) \), the necessary and sufficient condition for an intersection in interval 4 is 
\( \frac{\partial \Theta^N(\lambda, M)}{\partial \lambda} > \frac{\partial \Theta^P(\lambda, M)}{\partial \lambda} \), which is the strict version of Condition 3 given in (21). If there exists an interval 3 so that \( \Theta^N\left(\frac{M_{\mu2N}}{(1-\gamma)}, M\right) < \Theta^P\left(\frac{M_{\mu2N}}{(1-\gamma)}, M\right) \), the necessary and sufficient condition for an intersection in interval 4 is 
\( \frac{\partial \Theta^N(\lambda, M)}{\partial \lambda} > \frac{\partial \Theta^P(\lambda, M)}{\partial \lambda} \), which is the opposite of Condition 3 given in (21). The switching arrival rate \( \lambda_{s8} \) in interval 4 (if it exists) is defined by (73).

Combining the analysis in all intervals leads to the following result: If Condition 5 holds and Condition 3 either fails to hold or holds as an equality, \( \Theta^N(\lambda, M) \geq \Theta^P(\lambda, M) \) for all \( \lambda \geq 0 \). If Condition 5 fails to hold and Condition 3 holds, \( \Theta^N(\lambda, M) \geq \Theta^P(\lambda, M) \) for \( \lambda \leq \lambda_{s7} \), and \( \Theta^N(\lambda, M) < \Theta^P(\lambda, M) \) for \( \lambda > \lambda_{s7} \). If Condition 5 holds and Condition 3 holds as a strict inequality, \( \Theta^N(\lambda, M) \geq \Theta^P(\lambda, M) \) for \( \lambda \leq \lambda_{s8} \), and \( \Theta^N(\lambda, M) < \Theta^P(\lambda, M) \) for \( \lambda > \lambda_{s8} \). Finally, if Conditions 3 and 5 fail to hold, \( \Theta^N(\lambda, M) \geq \Theta^P(\lambda, M) \) for \( \lambda \leq \lambda_{s7} \) and \( \lambda \geq \lambda_{s8} \), and \( \Theta^N(\lambda, M) < \Theta^P(\lambda, M) \) for \( \lambda_{s7} < \lambda < \lambda_{s8} \). \( \square \)
E. An alternative implementation of Provider Triage at ED X

In this section, we investigate the structure of the proposed policy under an alternate implementation of PT where, instead of placing an additional provider to the triage step, one of the providers from the treatment area is moved to the triage step in the setting of ED X. In this case, having one less provider in the treatment area would reduce provider-to-bed ratio, and thus, would create a slowdown effect on the treatment step (in addition to the speedup effect discussed earlier). Therefore, this implementation would lead to smaller values of $\mu_{2P}$ than the original PT model we analyze with an additional provider in the triage area. However, because the total number of providers in the ED is fixed regardless of the triage method, there are no additional staffing costs.

In this section we use a method very similar to that in §5.2 to analyze this tradeoff.

We again use (18) to determine the triage method for given arrival rate and capacity level and approximate the objective function by (16). Determining the solution of (18) for the alternative implementation of PT in ED X also requires comparing the values of $\Theta^N$ and $\Theta^P$. Hence, the solution in this case is similar to those in Propositions 1 and 2 (Proposition 2 is provided in Appendix D). However, because $c_P - c_N = 0$, we cannot use the solutions in these results directly for all cases. In the rest of this section, we discuss the details of the solution for the alternative implementation of PT when the modified $\mu_{2P}$ satisfies the following mutually exclusive and exhaustive conditions:

(i) $\mu_{2P} \leq \mu_{2N}$ and the condition in part 2 of Proposition 2 holds, (ii) $\mu_{2P} \leq \mu_{2N}$ and the condition in part 1 of Proposition 2 holds, and (iii) $\mu_{2P} > \mu_{2N}$. In case (i), part 2 of Proposition 2 is still valid with $c_P - c_N = 0$.

When $\mu_{2P} \leq \mu_{2N}$ and the condition in part 1 of Proposition 2 holds, the structure of the solution is similar to that in part 1 of Proposition 2 but the exact solution is different. Before we present the details in Proposition 3 below, we list the following major differences between the solution for the alternative implementation and part 1 of Proposition 2. First, the profits from PT and NT overlap for small $\lambda$ values due to staffing costs being equal, i.e., $\Theta^N(\lambda) = \Theta^P(\lambda)$ for $\lambda \in [0, \bar{\lambda}]$, where

$$\bar{\lambda} = \frac{M \mu_{2N} \mu_{3N}}{\gamma \mu_{2N} + (1 - \gamma) \mu_{3N}},$$

unlike the original implementation of PT discussed in §5.2. Therefore, either triage method can be used for arrival rates less than $\bar{\lambda}$. Second, PT always outperforms NT for some interval of the arrival rate in all cases. This is because the profits from NT and PT are equal for all arrival rates smaller than $\bar{\lambda}$ and the slope of $\Theta^N$ decreases beyond this arrival rate while $\Theta^P$ keeps increasing at the same rate. In order to define the exact solution, we use Conditions 3 and 5 as given in (21) and (75).
Proposition 3. Assume that $\mu_{2P} \leq \mu_{2N}$, $\bar{\lambda} < \frac{M_{\mu_{2P}}}{(1-\gamma)}$ and $(c_P - c_N) = 0$. Under the alternative implementation of PT the optimal solution to (18) is given as follows.

1. $j^*(\lambda, M) = N$ or $P$ for $\lambda \in [0, \bar{\lambda}]$ and $j^*(\lambda, M) = P$ for $\lambda > \bar{\lambda}$ if Condition 5 holds and Condition 3 fails to hold.

2. $j^*(\lambda, M) = N$ or $P$ for $\lambda \in [0, \bar{\lambda}]$ and $\lambda \geq \Lambda_1$, and $j^*(\lambda, M) = P$ for $\lambda \in (\bar{\lambda}, \Lambda_1)$ if Condition 3 does not hold, where
   (a) $\Lambda_1$ is as given in (76) if in addition Condition 5 holds.
   (b) $\Lambda_1$ is as given in (24) if in addition Condition 5 fails to hold.

3. $j^*(\lambda, M) = N$ or $P$ for $\lambda \in [0, \bar{\lambda}]$ and $\lambda \in [\Lambda_1, \Lambda_2]$, and $j^*(\lambda, M) = P$ for $\lambda \in (\bar{\lambda}, \Lambda_1)$ and $\lambda > \Lambda_2$ if Condition 5 holds and Condition 3 holds as a strict inequality, where $\Lambda_1$ and $\Lambda_2$ are defined in (76) and (24).

When $\mu_{2P} > \mu_{2N}$, the solution under the alternative implementation of PT is similar to that provided in Proposition 1. The solution is significantly simpler because PT leads to an increase in the total capacity without increasing the staffing cost. In this case PT is definitely applied in one interval of the possible values of $\lambda$. In particular, PT and NT are equally profitable when arrival rate is below $\bar{\lambda}$, and PT outperforms NT when arrival rate is between $\bar{\lambda}$ and $\frac{M_{\mu_{2P}}}{(1-\gamma)}$. NT may or may not become more profitable as the arrival rate increases further depending on the abandonment cost difference between NT and PT periods.

Proof of Proposition 3: The proof for this case is very similar to that analyzed in Appendix D. First, it is trivial to show that $\Theta^N(\lambda, M) = \Theta^P(\lambda, M)$ for $\lambda \leq \frac{M_{\mu_{2N}}}{\gamma \mu_{2N} + (1-\gamma)\mu_{2N}}$, and $\Theta^P\left(\frac{M_{\mu_{2P}}}{(1-\gamma)}, M\right) \geq \Theta^N\left(\frac{M_{\mu_{2P}}}{(1-\gamma)}, M\right)$. Thus, a necessary and sufficient condition for an intersection in the interval $\left(\frac{M_{\mu_{2P}}}{(1-\gamma)}, \frac{M_{\mu_{2N}}}{(1-\gamma)}\right)$ becomes $\Theta^N\left(\frac{M_{\mu_{2N}}}{(1-\gamma)}, M\right) \geq \Theta^P\left(\frac{M_{\mu_{2N}}}{(1-\gamma)}, M\right)$, which is Condition 5 defined in (75).

The switching arrival rate in this interval (if it exists) is given by $\lambda_{s10}$ as given in (80).

If Condition 5 holds, $\Theta^N\left(\frac{M_{\mu_{2N}}}{(1-\gamma)}, M\right) \geq \Theta^P\left(\frac{M_{\mu_{2N}}}{(1-\gamma)}, M\right)$. In this case, another intersection when $\lambda > \frac{M_{\mu_{2N}}}{(1-\gamma)}$ exists if and only if Condition 3 as defined in (21) holds as a strict inequality. The switching arrival rate in this case, $\lambda_{s11}$, is given by (73). If Condition 5 fails to hold, then $\Theta^N\left(\frac{M_{\mu_{2N}}}{(1-\gamma)}, M\right) < \Theta^P\left(\frac{M_{\mu_{2N}}}{(1-\gamma)}, M\right)$. In this case, an intersection occurs at $\lambda_{s11}$ if and only if the opposite of Condition 3 as defined in (21) holds.

Then, if Condition 5 defined in (75) holds and Condition 3 in (21) holds strictly, $\Theta^P(\lambda, M) > \Theta^N(\lambda, M)$ for $\lambda \in \left(\frac{M_{\mu_{2N}}\mu_{2N}}{\gamma \mu_{2N} + (1-\gamma)\mu_{2N}}, \lambda_{s10}\right)$ and for $\lambda > \lambda_{s11}$, $\Theta^P(\lambda, M) \leq \Theta^N(\lambda, M)$ for $\lambda \in \left[0, \frac{M_{\mu_{2N}}\mu_{2N}}{\gamma \mu_{2N} + (1-\gamma)\mu_{2N}}\right]$ and $\lambda \in [\lambda_{s10}, \lambda_{s11}]$. If Condition 5 holds and Condition 3 fails to hold, $\Theta^P(\lambda, M) > \Theta^N(\lambda, M)$ for $\lambda \in \left(\frac{M_{\mu_{2N}}\mu_{2N}}{\gamma \mu_{2N} + (1-\gamma)\mu_{2N}}, \lambda_{s10}\right)$ and $\Theta^P(\lambda, M) \leq \Theta^N(\lambda, M)$ for $\lambda \in \left[0, \frac{M_{\mu_{2N}}\mu_{2N}}{\gamma \mu_{2N} + (1-\gamma)\mu_{2N}}\right]$. 

following conditions which will be used in Proposition 4. When
$$\Theta^P(\lambda, M) > \Theta^N(\lambda, M)$$ for
$$\lambda > \left( \frac{M\mu_{2N}\mu_{3N}}{\gamma\mu_{2N} + (1-\gamma)\mu_{3N}} \right)$$
and $$\lambda \geq \lambda_{s10}.$$ If Condition 5 fails to hold and Condition 3 holds, $$\Theta^P(\lambda, M) > \Theta^N(\lambda, M)$$ for $$\lambda \leq \left( \frac{M\mu_{2N}\mu_{3N}}{\gamma\mu_{2N} + (1-\gamma)\mu_{3N}} \right).$$ If Conditions 3 and 5 fail to hold, $$\Theta^P(\lambda, M) > \Theta^N(\lambda, M)$$ for $$\lambda \leq \left( \frac{M\mu_{2N}\mu_{3N}}{\gamma\mu_{2N} + (1-\gamma)\mu_{3N}}, \lambda_{s11} \right)$$ and $$\Theta^P(\lambda, M) \leq \Theta^N(\lambda, M)$$ for $$\lambda \leq \left( \frac{M\mu_{2N}\mu_{3N}}{\gamma\mu_{2N} + (1-\gamma)\mu_{3N}}, \lambda \geq \lambda_{s11}.$$

F. The Implementation of Proposed Solution in ED X for Limited PT Step Capacity

In this section, we study the triage decisions in the setting of ED X with limited PT step capacity. We examine the effect of PT step capacity limitation particularly when the abandonment cost before PT step, $$y_i,$$ is equal to the abandonment cost from treatment bed queue under NT method, $$w_{iN},$$ for $$i = 2, 3$$ because both terms reflect the cost of interventions applied at only the NT step. We also assume that PT step capacity (denoted by $$\tilde{\delta}$$), which is equal to $$\left( \frac{\delta w_{iN}}{(1-\gamma)\mu_{2N} + \gamma\mu_{3N}} \right),$$ is greater than the treatment step capacity under NT method, which is $$\left( \frac{M\mu_{2N}\mu_{3N}}{(1-\gamma)\mu_{2N} + \gamma\mu_{3N}} \right).$$ The exact structure of the solution depends on how $$\tilde{\delta}$$ compares to the treatment step capacity under NT when only treatment patients are treated, $$\frac{M\mu_{2N}}{(1-\gamma)},$$ and the treatment step capacity under PT method, $$\frac{M\mu_{2P}}{(1-\gamma)}.$$ We observe that when $$\tilde{\delta} > \frac{M\mu_{2N}}{(1-\gamma)},$$ the implications of Proposition 1 still hold when PT step capacity is considered but Condition 3 is ignored. In particular, when $$\tilde{\delta} \in \left( \frac{M\mu_{2N}}{(1-\gamma)}, \frac{M\mu_{2P}}{(1-\gamma)} \right),$$ the proposed solution is given by parts 1 (ignore the requirement for Condition 3), 2.(a) and 2.(b) of Proposition 1 where Condition 1 is as given in (19), Condition 2 is replaced by

$$\left( r_2 + w_{2N} \right) \left( (1-\gamma)\tilde{\delta} - M\mu_{2N} \right) + \left( r_3 + w_{3N} \right) \gamma\tilde{\delta} \geq c_P - c_N. \quad (84)$$

Thus, $$\Theta^N$$ and $$\Theta^P$$ can intersect at most once. Similarly, when $$\tilde{\delta} > \frac{M\mu_{2P}}{(1-\gamma)},$$ Proposition 1 yields the proposed solution by using Conditions 1 and 2 as given in (19) and (20) and by modifying Condition 3 in (21) as

$$\left( r_2 + w_{2P} \right) M\mu_{2P} - \left( r_2 + w_{2N} \right) M\mu_{2N} + \left( r_3 + w_{3N} \right) \gamma\tilde{\delta} + \left( w_{2N} - w_{2P} \right)(1-\gamma)\tilde{\delta} \geq c_P - c_N. \quad (85)$$

Hence, in this case, there can exist up to two threshold arrival rates. We observe that both for $$\tilde{\delta} \in \left( \frac{M\mu_{2N}}{(1-\gamma)}, \frac{M\mu_{2P}}{(1-\gamma)} \right)$$ and $$\tilde{\delta} > \frac{M\mu_{2P}}{(1-\gamma)},$$ the proposed triage method cannot change for arrival rates above $$\tilde{\delta}$$ because in this range $$\Theta^N$$ and $$\Theta^P$$ decrease at the same rate.

When $$\tilde{\delta} \in \left( \frac{M\mu_{2N}\mu_{3N}}{(1-\gamma)\mu_{2N} + \gamma\mu_{2P}}, \frac{M\mu_{2N}}{(1-\gamma)} \right),$$ however, the proposed policy is given by the next proposition, which also shows that $$\Theta^N(\lambda, M)$$ and $$\Theta^P(\lambda, M)$$ can intersect up to two times. We first define the following conditions which will be used in Proposition 4.

**Condition 6:** $$\left( r_3 + w_{3N} \right) \left( \gamma - \left( \frac{M - (1-\gamma)\tilde{\delta}}{\mu_{2N}} \right) \mu_{3N} \right) \geq c_P - c_N, \quad (86)$$

**Condition 7:** $$r_3 \gamma\tilde{\delta} + w_{3N} \left( \frac{M\mu_{2N}}{(1-\gamma)} - ((1-\gamma)r_2 + (1-\gamma)y_2 + \gamma y_3) \left( \frac{M\mu_{2N}}{(1-\gamma)} - \tilde{\delta} \right) \geq c_P - c_N. \quad (87)$$
Proposition 4. Assume that

$$\delta \in \left( \frac{M\mu_{2N}\mu_{3N}}{(1-\gamma)\mu_{3N}\gamma\mu_{2N}} \right).$$

Then

1. $j^*(\lambda, M) = N$ for all $\lambda \geq 0$, if Conditions 6 and 7 do not hold or hold 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 7 holds as a strict inequality, where
   (a) $\Lambda_1$ is as given in (22) if Condition 6 holds,
   (b) $\Lambda_1 = \frac{\mu_{2N}((1-\gamma)(y_2+r_2)+\gamma(y_3+r_3))\delta-(r_3+w_{3N})M\mu_{3N}-c_P+c_N)}{(1-\gamma)((r_2+y_2)\mu_{2N}-(r_3-w_{3N})\mu_{3N})-\gamma\mu_{2N}(w_{3N}-y_3)}$, (89)
   if Condition 6 does not hold.
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 6 strictly holds, and Condition 7 either does not hold or holds as an equality, where $\Lambda_1$ and $\Lambda_2$ are given by (22) and (89), respectively.

Proposition 4 shows that if (88) holds, there can exist up to two threshold arrival rates similar to Proposition 1. We also observe that for the cases when PT is implemented for arrival rates, $\lambda$, in the interval $(\Lambda_1, \Lambda_2)$, the maximum arrival rate at which PT is proposed to be implemented, $\Lambda_2$, is smaller when (88) holds compared to the ample PT capacity case in Proposition 1. Hence, due to higher costs of abandonments before PT step, the proposed policy recommends switching back to NT method at smaller arrival rates than that in Proposition 1.

Proof of Proposition 4: When $\frac{M\mu_{2N}\mu_{3N}}{(1-\gamma)\mu_{3N}\gamma\mu_{2N}} < \delta \leq \frac{M\mu_{2N}}{(1-\gamma)\mu_{3N}+\gamma\mu_{2N}}$, the analysis when $\lambda \leq \frac{M\mu_{2N}\mu_{3N}}{(1-\gamma)\mu_{3N}+\gamma\mu_{2N}}$ (that is, interval 1 in the proof of Proposition 1) is the same as in Appendix B. We additionally define the following intervals of $\lambda$. Interval 2 is defined by $\frac{M\mu_{2N}\mu_{3N}}{(1-\gamma)\mu_{3N}+\gamma\mu_{2N}} < \lambda \leq \bar{\delta}$. In this interval, $a_{2N}(\lambda, M) = 0, a_{3N}(\lambda, M) > 0, a_{iP}(\lambda, M) = 0$ and $b_{iN}(\lambda, M) = 0$ for $i = 2, 3$. Interval 3 is defined by $\bar{\delta} < \lambda \leq \frac{M\mu_{2N}}{(1-\gamma)}$. In interval 3, we have $a_{2N}(\lambda, M) = 0, a_{3N}(\lambda, M) > 0, a_{iP}(\lambda, M) = 0$ and $b_{iN}(\lambda, M) = 0$ for $i = 2, 3$. Finally, interval 4 is defined by $\lambda > \frac{M\mu_{2N}}{(1-\gamma)}$, which leads to $a_{iN}(\lambda, M) > 0, a_{iP}(\lambda, M) = 0$ and $b_{iN}(\lambda, M) > 0$ for $i = 2, 3$.

In interval 2, $\Theta^N(\lambda, M)$ and $\Theta^P(\lambda, M)$ are as given in (59) and (60). Hence, by (58), a necessary and sufficient condition for an intersection in interval 2 is

$$\Theta^N(\delta, M) = r_2(1-\gamma)\delta + r_3\left(M - \frac{(1-\gamma)\delta}{\mu_{2N}}\right)\mu_{3N} - w_{3N}\left(\gamma\delta - \left(M - \frac{(1-\gamma)\delta}{\mu_{2N}}\right)\mu_{3N}\right) - c_N$$
$$\leq \Theta^P(\delta, M) = r_2(1-\gamma)\delta + r_3\gamma\delta - c_P,$$
which is equivalent to (86). Thus, if and only if (86) holds, there exists a switching arrival rate \( \lambda_{s1} \), i.e., \( \Theta^{N}(\lambda_{s1}, M) = \Theta^{P}(\lambda_{s1}, M) \), that is given by (63).

In interval 3, \( \Theta^{N}(\lambda, M) \) is given by (59), and \( \Theta^{P}(\lambda, M) \) is given by

\[
\Theta^{P}(\lambda, M) = -y_{2}(1-\gamma)(\lambda - \bar{\delta}) - y_{3}\gamma(\lambda - \bar{\delta}) + r_{2}(1-\gamma)\bar{\delta} + r_{3}\gamma\bar{\delta} - c_{P}. \tag{91}
\]

Thus, if (86), and hence (90), does not hold, there exists an intersection in interval 3 if and only if

\[
\Theta^{N}\left(\frac{M\mu_{2N}}{1-\gamma}, M\right) = r_{2}M\mu_{2N} - w_{3N}\frac{\gamma M\mu_{2N}}{(1-\gamma)} - c_{N}
\leq \Theta^{P}\left(\frac{M\mu_{2N}}{1-\gamma}, M\right) = -y_{2}(1-\gamma)\left(\frac{M\mu_{2N}}{1-\gamma} - \bar{\delta}\right) - y_{3}\gamma\left(\frac{M\mu_{2N}}{1-\gamma} - \bar{\delta}\right) + r_{2}(1-\gamma)\bar{\delta} + r_{3}\gamma\bar{\delta} - c_{P}, \tag{92}
\]

which is equivalent to (87).

In interval 4, \( \Theta^{N}(\lambda, M) \) is given by (69) and \( \Theta^{P}(\lambda, M) \) is given by (91), which implies that

\[
\frac{\partial \Theta^{N}(\lambda, M)}{\partial \lambda} = -(1-\gamma)w_{2N} - \gamma w_{3N} = \frac{\partial \Theta^{P}(\lambda, M)}{\partial \lambda} = -(1-\gamma)y_{2} - \gamma y_{3}. \tag{93}
\]

Hence, the proposed solution cannot change in interval 4, and the conditions in (86) and (87) suffice to define the proposed solution.

Combining the analysis of all intervals, if (86), and hence (90), does not hold or holds as an equality, and (87), and hence (92), does not hold or holds as an equality, \( \Theta^{N}(\lambda, M) \geq \Theta^{P}(\lambda, M) \) for all \( \lambda \geq 0 \), which proves part 1 of Proposition 4. If (86) holds and (87) holds as a strict inequality, \( \Theta^{P}(\lambda, M) > \Theta^{N}(\lambda, M) \) holds if and only if \( \lambda > \Lambda_{1} \), where \( \Lambda_{1} \) is as given in (22), which proves part 2.(a) of the proposition. Similarly, if (86) does not hold and (87) holds as a strict inequality, \( \Theta^{P}(\lambda, M) > \Theta^{N}(\lambda, M) \) holds if and only if \( \lambda > \Lambda_{1} \), where \( \Lambda_{1} \) is as given in (89), which proves part 2.(b) of the proposition. Finally, if (86) holds as a strict inequality and (87) either does not hold or holds as an equality, \( \Theta^{P}(\lambda, M) > \Theta^{N}(\lambda, M) \) holds if and only if \( \lambda \in (\Lambda_{1}, \Lambda_{2}) \), where \( \Lambda_{1} \) and \( \Lambda_{2} \) are as given in (22) and (89), respectively, which proves part 3 of Proposition 4. \( \square \)

G. Proof of Theorem 1

G.1. Definitions

Before we can prove the result, we need to introduce additional notation. We denote the patient type as \((i, j)\), where \( i = 2 \) and \( i = 3 \) denote treatment and triage patients, respectively. The triage method that has been used at the arrival time of the patient is denoted as \( j \), \( j \in \{N, P\} \). So, the patient type \((i, j) \in S = \{(2, N), (2, P), (3, N)\}\). Let \( A_{ij}(t) \) denote the number of arrivals to the queue of the treatment stage, \( D_{ij}(t) \) denote the number of discharged patients, \( R_{ij}(t) \) denote the number of abandoned patients, and \( K_{ij}(t) \) denote the number of patients transferred into the
treatment beds for type \((i,j)\) patients by time \(t\). Let \(Z_{ij}(t)\) and \(Q_{ij}(t)\) denote the number of type \((i,j)\) patients in the treatment beds and in the queue at time \(t\), respectively.

Then, for any policy and for \((i,j) \in S\) we have

\[
Z_{ij}(t) = Z_{ij}(0) + K_{ij}(t) - D_{ij}(t),
\]

\[
Q_{ij}(t) = Q_{ij}(0) + A_{ij}(t) - R_{ij}(t) - K_{ij}(t),
\]

\[
M \geq Z_{ij}(t) \geq 0,
\]

\[
Q_{ij}(t) \geq 0.
\]

Let \(\theta_{ij}\) and \(\mu_{ij}\) denote the abandonment and service rates, respectively, for each patient type \((i,j) \in S\). We denote by \(\Lambda_i(t)\) the arrival rate of type \(i\) patients at time \(t\). Due to our assumption that interarrival times, treatment times and abandonment times are exponential, we have

\[
D_{ij}(t) = S_{ij} \left( \mu_{ij} \int_0^t Z_{ij}(s) \, ds \right) \quad \text{and} \quad R_{ij}(t) = G_{ij} \left( \theta_{ij} \int_0^t Q_{ij}(s) \, ds \right),
\]

for \((i,j) \in S\) under any Markovian policy where \(S_{ij}\) and \(G_{ij}\) are rate 1 Poisson processes.

Let \(A_i(t)\) denote the number of type \(i\) patients who completed the initial triage step by time \(t\). Hence, if the PT is applied at time \(t\), these arriving patients would be arriving at the triage process by the provider. Then

\[
A_i(t) = F_i \left( \int_0^t \Lambda_i(s) \, ds \right),
\]

where \(F_i\) is a rate 1 Poisson process, \(i = 2, 3\). For simplicity, we assume that the time for the provider triage is fixed and we denote it by \(\tau_P\). Then, \(A_{ij}(t)\) for type \((i,j) \in S\) patient is given by

\[
A_{2P}(t) = \int_0^{t-\tau_P} \mathbb{1} \{ \pi(s) = P \} \, dA_2(s), \quad \text{for } t \geq \tau_P,
\]

\[
A_{2N}(t) = \int_0^t \mathbb{1} \{ \pi(s) = N \} \, dA_2(s), \quad A_{3N}(t) = \int_0^t \mathbb{1} \{ \pi(s) = N \} \, dA_3(s).
\]

We assume for simplicity that there are no patients in provider triage at time zero. The total queue length, \(Q\), is given by

\[
Q(t) = \sum_{(i,j) \in S} Q_{ij}(t).
\]

Also, because all triage patients leave the system after triage step during PT, we have

\[
D_{3P}(t) = \int_0^{t-\tau_P} \mathbb{1} \{ \pi(s) = P \} \, dA_3(s), \quad \text{for } t \geq \tau_P.
\]
G.2. Proof of Theorem 1 Part (i)

We consider a sequence of ED’s indexed by \( n \) that satisfy (25)-(29) and (33). We append superscript “\( n \)” to all queueing processes defined in Appendix G.1 associated with the \( n \)th ED. Fix a sequence of admissible policies \( \pi^n \). We prove the following compact containment condition in §EC1.3, see Remark 4.2 in Dai and Tezcan (2011). Let \( \tilde{Q}^n(t) = \frac{Q^n(t)}{n} \), where \( Q(t) \) is defined as in (95) and (102).

**Lemma 1.** Under assumptions (25)–(29) and (33), there exists a constant \( B \) such that

\[
\lim_{n \to \infty} P \left( \int_0^T \tilde{Q}^n_i(k^n s) \, ds > B \right) = 0 \tag{103}
\]

and

\[
\lim_{C \to \infty} \lim_{n \to \infty} P \left( \|\tilde{Q}^n(t)\|_{k^nT} > C \right) = 0, \tag{104}
\]

under any Markovian policy for any \( T > 0 \).

Recall that \( F_i, S_{ij} \) and \( G_{ij} \) are Poisson processes with rate 1 and \( A_{ij}, D_{ij} \) and \( R_{ij} \) are as defined in (98)–(101), for \( i = 2, 3 \) and \( (i, j) \in S \). We define

\[
\bar{F}^n_i(t) = \frac{F_i(k^n t)}{k^n n}, \quad \bar{S}^n_{ij}(t) = \frac{S_{ij}(k^n t)}{k^n n}, \quad \bar{G}^n_{ij}(t) = \frac{G_{ij}(k^n t)}{k^n n}, i = 2, 3, (i, j) \in S.
\]

Fix \( T > 0 \) and choose \( B \) such that (103) holds. Also fix \( \delta > 0 \), set \( C = 2c_\lambda T^2 \vee B \) and let

\[
\Xi^n(\delta) = \left\{ \sup_{0 \leq t \leq C} \left| \bar{S}^n_{ij}(t) - \bar{S}^n_{ij}(t) \right| < \delta; (i, j) \in S \right\} \cup \left\{ \sup_{0 \leq t \leq C} \left| \bar{G}^n_{ij}(\theta_{ij} t) - \bar{G}^n_{ij}(\theta_{ij} t) \right| < \delta; (i, j) \in S \right\}.
\]

By (103), as in Atar et al. (2010) and Pinker and Tezcan (2013), we have

\[
P\Xi^n(\delta) \to 1 \quad \text{as} \quad n \to \infty. \tag{106}
\]

We additionally define the following:

\[
U^n_{ij}(t) = \int_0^t Z^n_{ij}(s) \, ds, \quad L^n_{ij}(t) = \int_0^t Q^n_{ij}(s) \, ds, \tag{107}
\]

\[
\tilde{U}^n_{ij}(t) = \frac{U^n_{ij}(k^n t)}{k^n n}, \quad \tilde{L}^n_{ij}(t) = \frac{L^n_{ij}(k^n t)}{k^n n}, \quad \tilde{A}^n_{ij}(t) = \frac{A^n_{ij}(k^n t)}{k^n n}, \quad \tilde{Q}^n_{ij}(t) = \frac{Q^n_{ij}(k^n t)}{k^n n}, \tag{108}
\]

\[
\tilde{Z}^n_{ij}(t) = \frac{Z^n_{ij}(k^n t)}{k^n n}, \quad \tilde{K}^n_{ij}(t) = \frac{K^n_{ij}(k^n t)}{k^n n}, \quad \tilde{D}^n_{ij}(t) = \frac{D^n_{ij}(k^n t)}{k^n n}, \quad \tilde{R}^n_{ij}(t) = \frac{R^n_{ij}(k^n t)}{k^n n}. \tag{109}
\]

By (94)–(97), for \( (i, j) \in S \) we have

\[
\tilde{Z}^n_{ij}(t) = \tilde{Z}^n_{ij}(0) + \tilde{K}^n_{ij}(t) - \tilde{D}^n_{ij}(t), \tag{110}
\]
\begin{equation}
\tilde{Q}_{ij}(t) = \tilde{Q}_{ij}(0) + \tilde{A}_{ij}(t) - \tilde{R}_{ij}(t) - \tilde{K}_{ij}(t), \tag{111}
\end{equation}

\begin{equation}
0 \leq \sum_{(i,j) \in S} \tilde{Z}_{ij}(t) \leq \frac{M_x}{k^n}, \tag{112}
\end{equation}

\begin{equation}
\tilde{Q}_{ij}(t) \geq 0. \tag{113}
\end{equation}

Obviously

\[ \sup_{0 \leq t \leq T} \tilde{Z}_{ij}(t) \to 0 \text{ as } n \to \infty \text{ a.s.}, \]

by (112). From Lemma 1, we can assume, without loss of generality, that on \( \Xi^n(\delta) \)

\[ \sup_{0 \leq t \leq T} \tilde{Q}_{ij}(t) \to 0 \text{ as } n \to \infty. \]

On \( \Xi^n(\delta) \), by (25) and Lemma 1, we have for \((i,j) \in S\)

\begin{equation}
\sup_{0 \leq t \leq T} \left| \tilde{A}_{ij}(t) - \int_0^t 1 \{ \pi^n(k^n s) = j \} \Lambda_i(s) ds \right| < \delta + \frac{2cA TP}{k^n}, \tag{114}
\end{equation}

\begin{equation}
\sup_{0 \leq t \leq T} \left| \tilde{D}_{ij}(t) - \mu_{ij} \tilde{U}_{ij}(t) \right| < \delta, \tag{115}
\end{equation}

\begin{equation}
\sup_{0 \leq t \leq T} \left| \tilde{R}_{ij}(t) - \theta_{ij} \tilde{L}_{ij}(t) \right| < \delta. \tag{115}
\end{equation}

Also

\begin{equation}
\sup_{0 \leq t \leq T} \left| \tilde{D}_{3P}(t) - \int_0^t 1 \{ \pi^n(k^n s) = P \} \Lambda_3(s) ds \right| < \delta + \frac{2cA TP}{k^n}. \tag{116}
\end{equation}

By (110)–(116) and (27), for all \( 0 \leq t \leq T \) and \((i,j) \in S\)

\begin{equation}
|\tilde{K}_{ij}(t) - \mu_{ij} \tilde{U}_{ij}(t)| \leq 3\delta, \tag{117}
\end{equation}

\begin{equation}
\left| \int_0^t \Lambda_i(s) 1 \{ \pi^n(k^n s) = j \} ds - \theta_{ij} \tilde{L}_{ij}(t) - \tilde{K}_{ij}(t) \right| \leq 4\delta, \tag{118}
\end{equation}

\begin{equation}
\sum_{(i,j) \in S} \tilde{U}_{ij}(t) \leq \int_0^T M(t) dt. \tag{119}
\end{equation}

By (4)–(14), (26) and (117)–(119), we have

\begin{equation}
\sum_{(i,j) \in S} \left( r_i \tilde{D}_{ij}(T) - \omega_{ij} \tilde{R}_{ij}(T) \right) - \frac{1}{k^n n} \sum_{j \in \{N, P\}} c_j^p \int_0^{k^n T} 1 \{ \pi^n(s) = j \} ds \leq \int_0^T \Theta^{n^*}(t) (\Lambda(t)) dt + b\delta, \tag{120}
\end{equation}

for some \( b > 0 \) on \( \Xi^n(\delta) \). By (18) and (106), (120) yields (34) because \( \delta > 0 \) is arbitrary. □
G.3. Proof of Theorem 1 Part (ii)

We next analyze our proposed solution under the asymptotic regime explained in §5.4. In Proposition 1 we showed that, under the threshold policy, triage method either does not change as the arrival rate varies or there exists at most two threshold arrival rates at which triage method changes. For the purposes of asymptotic analysis, we focus on Case 2 in Proposition 1; that is, there is a single threshold arrival rate \( \Lambda_1 \). Our analysis can easily be extended to other cases with zero or two threshold arrival rates by following the arguments in the proof of Theorem 2 below.

Recall that \( \pi^*(t) \) and \( \pi^{*,n}(t) \) are the triage methods at time \( t \) suggested by the threshold policy for the original and \( n \)th systems as defined in (31) and (32), respectively. Let \( z_{ij}(\lambda, M) \) and \( q_{ij}(\lambda, M) \) denote the number of type \( i \) patients in treatment beds and in queue when triage method \( j \) is applied under fluid approximations, where \( \lambda \) is the arrival rate and \( M \) is the number of beds, \( i \in \{2,3\} \) and \( j \in \{N,P\} \). Thus, \( z_{ij}(\lambda, M) \) and \( q_{ij}(\lambda, M) \) are defined as follows:

\[
z_{ij}(\lambda, M) = \frac{d_{ij}(\lambda, M)}{\mu_{ij}}, \quad q_{ij}(\lambda, M) = \frac{a_{ij}(\lambda, M)}{\theta_{ij}},
\]

where \( d_{ij}(\lambda, M) \) and \( a_{ij}(\lambda, M) \) are defined in (4)–(14), \( (i,j) \in S \). We additionally define \( z_{ij}^\pi \) and \( q_{ij}^\pi \) as

\[
z_{ij}^\pi(\lambda, M) = \begin{cases} z_{i\pi}(\lambda, M), & \text{if } j = \pi \\ 0, & \text{otherwise} \end{cases}, \quad q_{ij}^\pi(\lambda, M) = \begin{cases} q_{i\pi}(\lambda, M), & \text{if } j = \pi \\ 0, & \text{otherwise} \end{cases},
\]

where \( z_{i\pi} \) and \( q_{i\pi} \) are defined in (121) and \( (i,\pi) \in S \). Next we define for \( (i,j) \in S \)

\[
\tilde{Z}_{ij}^n(t) = \frac{Z_{ij}^n(k^nt)}{n} \quad \text{and} \quad \tilde{Q}_{ij}^n(t) = \frac{Q_{ij}^n(k^nt)}{n}.
\]

Let \( \mathcal{D} \) denote the set of points where \( \pi^* \) is discontinuous and recall that by Assumption 1 it is finite.

**Theorem 2.** Consider a sequence of ED’s indexed by \( n \) that satisfy (25)–(29), (33) and Assumption 1. Under the threshold policy defined as in (32), for any \( T > 0 \) and \( \epsilon > 0 \)

\[
\lim_{n \to \infty} P\{ \sup_{t \in (0,T) \setminus \mathcal{D}} |\tilde{Z}_{ij}^n(t) - z_{ij}^{\pi^*(t)}(\Lambda(t), M(t))| > \epsilon \} = 0,
\]

\[
\lim_{n \to \infty} P\{ \sup_{t \in (0,T) \setminus \mathcal{D}} |\tilde{Q}_{ij}^n(t) - q_{ij}^{\pi^*(t)}(\Lambda(t), M(t))| > \epsilon \} = 0,
\]

where \( (i,j) \in S \) and \( \pi^* \) is defined in (31).

We now prove part (ii) of Theorem 1. Consider a sequence of ED’s that satisfy (25)–(29) and Assumption 1. We fix \( \delta > 0 \), define \( \Xi^n(\delta) \) as in (105) (also recall (106)) and consider the scaling
in (108) and (109). We note that (114) and (115) hold under threshold policy. Thus, we have on 
\[ \Xi_n(\delta) \]
\[ \sup_{t \in (0,T) \setminus \mathcal{D}} \left| \tilde{Y}_{ij}^n(t) - \mu_{ij} \int_0^t \tilde{Z}_{ij}^n(s) \, ds \right| < \delta, (i,j) \in S, \]
\[ \sup_{t \in (0,T) \setminus \mathcal{D}} \left| \tilde{R}_{ij}^n(t) - \theta_{ij} \int_0^t \tilde{Q}_{ij}^n(s) \, ds \right| < \delta, (i,j) \in S, \]

By (30), (32), (106) and Theorem 2, under threshold policy

\[ \tilde{Y}_{ij}^n(T) \to \mu_{ij} \int_0^T \pi_{ij}^*(s)(\Lambda(s), M(s)) \, ds, (i,j) \in S, \] (124)
\[ \tilde{R}_{ij}^n(T) \to \theta_{ij} \int_0^T q_{ij}^*(s)(\Lambda(s), M(s)) \, ds, (i,j) \in S, \] (125)
in probability as \( n \to \infty \). The result follows from (25), (26), (31), (32), (124), (125), and dominated convergence theorem.

The proof of Theorem 2 is based on analyzing the limits of processes \( Z_{ij}^n \)'s and \( Q_{ij}^n \)'s in the asymptotic regime described in §G.3 under the proposed policy. We show that the limiting processes satisfy a set of equations that are similar to many-server fluid model equations (see Dai and Tezcan (2011)). These equations are known as hydrodynamic equations in the literature (see Bramson (1998) and Dai and Tezcan (2011)), and in the current asymptotic regime they agree with the more traditional many-server fluid model equations. Our approach in this proof is similar to that of Theorem 3 in Pinker and Tezcan (2013) and Besbes and Maglaras (2009). However, there are significant differences. First, the policy we study here changes how the arrivals are routed to the different parts of the system, whereas Pinker and Tezcan (2013) and Besbes and Maglaras (2009) only studied dispatching policies. Also, we consider systems with unlimited queueing space, while Pinker and Tezcan (2013) studied loss systems, and Besbes and Maglaras (2009) considered a system in which customers do not join the system if the queue length reaches a certain level. In addition, we need to deal with discontinuities in our policy. The details of the proof is presented in the electronic companion, see §EC1.
H. Results of Sensitivity Analysis

Figure 7 Change in $\Lambda_1$ vs. $\mu_{2N}$, $\mu_{3N}$, $\gamma$ and $M$ when $w_{3N}$ is $0$, $100$ and $500$. 