DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY

Vertical Patient Streaming in Emergency Departments

Citation

Feizi, Arshya, Agni Orfanoudaki, Soroush Saghafian, and Nicole Hodgson. "Vertical Patient Streaming in Emergency Departments." HKS Faculty Research Working Paper Series RWP23-014, May 2023.

Link

https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37375415

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles (OAP), as set forth at
https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

Accessibility

https://accessibility.huit.harvard.edu/digital-accessibility-policy

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story

HARVARD Kennedy School
JOHN F. KENNEDY SCHOOL OF GOVERNMENT

Vertical Patient Streaming in Emergency Departments

Faculty Research Working Paper Series

Arshya Feizi
Harvard University

Agni Orfanoudaki
University of Oxford

Soroush Saghafian
Harvard Kennedy School

Nicole Hodgson
Mayo Clinic

May 2023

RWP23-014

Visit the **HKS Faculty Research Working Paper Series** at: https://ken.sc/faculty-research-working-paper-series

The views expressed in the **HKS Faculty Research Working Paper Series** are those of the author(s) and do not necessarily reflect those of the John F. Kennedy School of Government or of Harvard University. Faculty Research Working Papers have not undergone formal review and approval. Such papers are included in this series to elicit feedback and to encourage debate on important public policy challenges. Copyright belongs to the author(s). Papers may be downloaded for personal use only.

Vertical Patient Streaming in Emergency Departments

Arshya Feizi
Harvard University, afeizi@hks.harvard.edu

Agni Orfanoudaki
University of Oxford, agni.orfanoudaki@sbs.ox.ac.uk

Soroush Saghafian
Harvard University, soroush_saghafian@hks.harvard.edu

Nicole Hodgson
Mayo Clinic, Hodgson.Nicole@mayo.edu

Addressing hospital emergency department (ED) overcrowding is a critical challenge for many healthcare systems worldwide. Many hospitals (including our partner hospital) have been experimenting with innovative patient flow designs to address this challenge. A promising new design is to separate patients who can be served vertically (e.g., on a regular chair as opposed to horizontally on an ED bed) and route them to a different area termed the Vertical Processing Pathway (VPP) unit. While this can potentially increase operational efficiency by removing the burden caused by a main ED bottleneck—lack of bed availability—it can degrade performance if patients that are routed to the VPP unit need to be sent back to be served in an ED bed, or if some patients that could have been served in the VPP unit end up occupying an ED bed. Successful implementation of this design, thus, significantly depends on understanding which patients should be routed to the VPP unit and when.

To assist our partner hospital and other EDs, we develop a machine learning model trained on large-scale data capable of providing a personalized risk score for each arriving patient on whether or not they will eventually need an ED bed. We then feed these risk scores to an analytical model of patient flow to characterize the optimal protocol for utilizing the VPP unit. We find that the optimal protocol depends not only on the predicted risk scores but also on the machine learning model's accuracy as well as some of the main ED characteristics (e.g., patient arrival intensity and congestion level). To gain deeper insights, we make use of simulation analyses calibrated with hospital data and compare the performance of our recommended VPP-based patient streaming design with more traditional ED flow approaches such as "fast track" or "physician in triage." Our results suggest that following the VPP design under our recommended protocol can bring several advantages to EDs, allowing them to significantly improve their operations.

Key words: Emergency Department, Machine Learning, Operational Efficiency; Vertical Processing; Patient Flow

1. Introduction

Emergency Department (ED) overcrowding has been reported as a major issue of many healthcare systems throughout the past two decades (Schafermeyer and Asplin 2003). The COVID-19 pandemic further aggravated this problem, resulting in significant increases in

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

emergency medicine patient volumes and additional delays and overcrowding to already strained EDs. ED overcrowding is often exacerbated due to the "bed-block" problem, where ED patients who need to be transferred to hospital inpatient units end up occupying ED beds for long hours, mainly because of lack of bed availability in inpatient units (Saghafian et al. 2023). To better understand this phenomenon, it is worth noting that, between 2010 and 2017, the number of total ED visits increased from 128.97 million to 144.82 million in the U.S. (Lane et al. 2020, Moore and Liang 2020). Despite this, the hospital supply of ED beds remained relatively stable (Schafermeyer and Asplin 2003, Venkatesh et al. 2021).

The combination of the aforementioned trends has escalated ED overcrowding to crisis levels in many hospitals. Strained ED departments lead to excessive wait times for ED treatment, compromising patient safety, not only within the ED but also throughout the entire healthcare system (Di Somma et al. 2015, Wong et al. 2010). ED overcrowding has been shown to cause delays in diagnosis and treatment, leading to poor patient outcomes and quality of care (Association et al. 2002). The repercussions become even more prominent in the cases of critically ill patients who remain an especially vulnerable population, including those with an acute coronary syndrome, surgical emergencies, stroke, and septic shock (Cowan and Trzeciak 2004, Derlet and Richards 2002). For those patients who are not critically ill, overcrowding often leads to extremely long waiting times, fueling patient dissatisfaction and walkouts, which can pose a threat to long-term, high-quality medical care (Cowan and Trzeciak 2004). In addition, ED overcrowding causes significant stress and overburden to physicians, increasing the risk of medical errors (Rondeau et al. 2005). It can also lead to ambulance diversion and threaten disaster preparedness (Olshaker and Rathlev 2006).

A direct way of addressing ED congestion is to deploy additional resources (e.g., physicians, nurses, beds, or testing capacity). However, this remains a substantially expensive option that is constrained by the physical space and financial resources. An alternative method to address the overcrowding problem is to make use of advanced technology such as telemedical triage, which allows EDs to offload some of their tasks to physicians that are serving patients in a different hospital (Saghafian et al. 2018). Alternatively, the ED can choose to "close the doors" through ambulance diversion. However, the Emergency Medical Treatment and Labor Act (EMTALA) allows this approach only in the case of an internal disaster. Under regular conditions, EMTALA mandates EDs to serve all patients

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

who present to the facility, regardless of their insurance or financial status (Fields et al. 2001). Furthermore, ambulance diversion is not legal in some states even under congested conditions. Therefore, an attractive approach to address these problems in EDs has been to optimize patient flow processes, which may not require a significant investment in technology or additional resources.

Several studies have demonstrated that optimizing the ED patient flow process can result in significant improvements (Saghafian et al. 2015). Patient streaming, for example, is a well-established ED flow design (Saghafian et al. 2012), which in its basic format, improves performance by separating and routing patients into distinct streams according to their anticipated disposition. In contrast to pooling, where all types of patients are treated by the same resources, this type of streaming can lead to improved system efficiency by separating resources for patients that are predicted to be discharged home from ED and those that might require hospital admission post-ED service (Saghafian et al. 2012). Streaming can also be implemented in various other ways, including streaming based on medical complexity (Saghafian et al. 2014), by using a dedicated Fast-Track (FT), implementing a Physician In Triage (PIT) approach, or making use of a Vertical Processing Pathway (VPP) unit (Hodgson et al. 2023). While FT and PIT approaches have been widely adopted by EDs across the world, VPP still remains a hybrid, ad-hoc design. In particular, it has been primarily proposed and adopted by our partner hospital, the Mayo Clinic (Hodgson et al. 2023). At the time of our analysis, our partner hospital prepares the opening of a new building for their ED, where the administration plans to make use of a fully operational VPP unit. Our aim is to assist the administration at our partner hospital by studying and recommending the best ways of utilizing the VPP unit. We also draw conclusions from our study to help other hospitals in deciding whether and how a VPP-based flow design should be implemented.

To these ends, we use a combination of analytical, Machine Learning (ML), and simulation models, and we address the following questions:

- *Can an ML model be developed to accurately predict whether an arriving ED patient can be served in the VPP unit without eventually requiring an ED bed?*
- *Given predictions from an ML model, what patients should be prioritized for routing to the VPP unit given the characteristics of the ED? What routing protocol optimizes the performance under the VPP patient flow design?*

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

- *For what hospitals does the VPP-based flow design outperform the FT and/or PIT designs?*

1.1. Introducing the FT, VPP, and PIT Flow Designs

Prior to addressing our research questions, it is helpful to first introduce and discuss the key differences between FT, VPP, and PIT patient flow designs. Figure 1 illustrates the patient flow in the ED under these three forms of patient streaming. Table 1 summarizes their main differences in terms of who triages patients, how low-complexity patients are initially determined (i.e., the selection criteria), and whether the selected low-complexity patients are separated from the rest of the patients and assigned to a dedicated queue.

In an FT model, arriving patients are first triaged and assigned an Emergency Severity Index (ESI) from 1 (most urgent) to 5 (least urgent) by a triage nurse. Patients with an ESI greater than 3 are routed to a separate dedicated queue to be treated in a section of the ED called the FT. In some hospitals, FT providers comprise nurse practitioners or physician assistants dedicated to managing patients in that section of the ED. The main idea of the FT design is to avoid having low acuity patients (who often have shorter "processing times") wait behind high acuity ones.

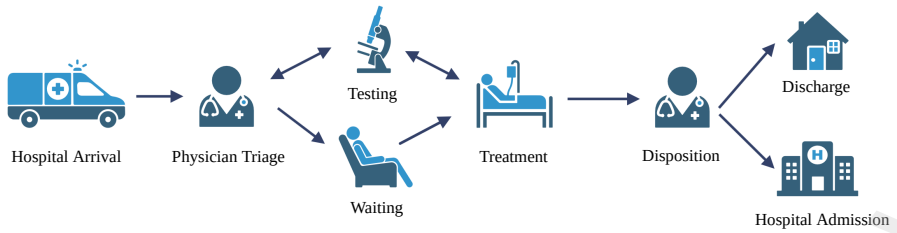
In a PIT model, as the name suggests, a medical provider licensed to order tests and perform the treatment (e.g., physician or advanced practice provider) is assigned to the triage stage working alongside a registered nurse (Franklin et al. 2021). PIT systems essentially provide more flexibility and a higher degree of responsibility to the stage of triage, leveraging medical experts with more advanced training. In addition to assigning an ESI score, (a) ED tests can be initialized during triage, and (b) patients who do not need sophisticated ED care are identified and discharged. Thus, triage providers have the discretion to disposition patients directly. The PIT model has various benefits and drawbacks, as discussed in the literature (Franklin et al. 2021) and has also been implemented in various formats over time (Traub et al. 2015, 2016). The operational success of such systems depends on local contextual factors, and thus, mixed empirical results have been reported in the literature (Benabbas et al. 2020).

Finally, in the VPP model, patients are triaged by a nurse and assigned an ESI level and are then asked to wait to be seen in the main treatment area. However, a physician

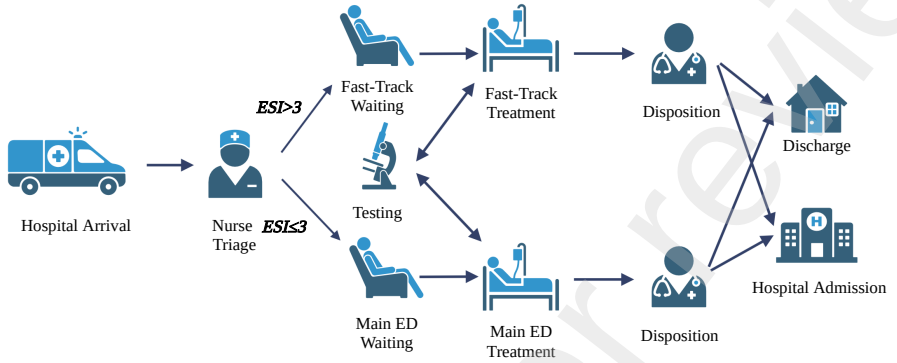
[1] $\gamma \in [0,1]$: unknown medical complexity level; $\tau \in [0,1]$: classification threshold for the VPP.

Preprint not peer reviewed

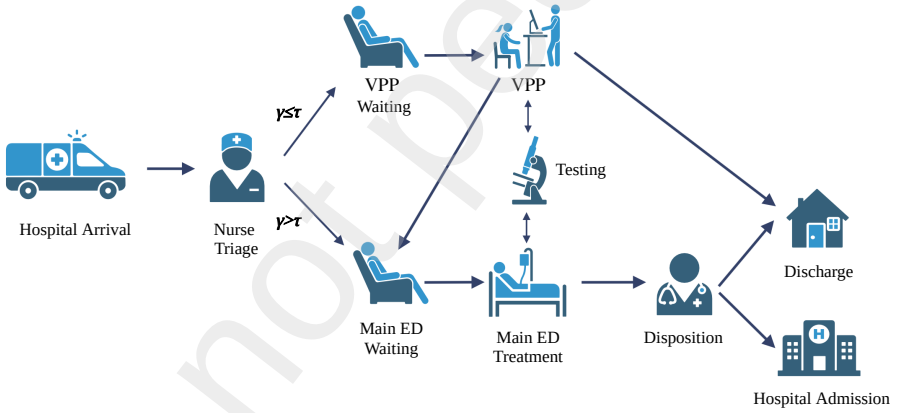
This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161



(a) Physician in Triage (PIT) design



(b) Fast-Track (FT) design



(c) Vertical Processing Pathway (VPP) design[1]

Figure 1 ED patient flow designs.

Model	Triage Staff	Assumed Low-Complexity	Dedicated Queue	Test Ordering
FT	RN	ESI > 3	Yes	Yes
PIT	PA/MD	All patients	No	Yes
VPP	RN	Doctor's discretion	No	Yes

Table 1 A Comparison Between FT, VPP, and PIT.

Notes. RN: Registered Nurse, PA: Physician Assistant, MD: Medical Doctor

can evaluate a waiting patient's triage data and guess whether the patient is suitable for treatment and discharge from the VPP unit after a quick visit (e.g., 15–20 minutes)

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

without needing an ED bed. If the patient is evaluated in the VPP, but the physician realizes that she needs to be treated on an ED bed, the patient is sent back to the waiting area and is asked to wait until a bed becomes available.[2] An important difference between the VPP model and the PIT model is that under the latter, every patient is seen by a physician during triage, and the ED assigns resources to the triage stage to reflect this. In contrast, under the former, only a portion of patients are routed to the VPP because it is not physically equipped or designed to handle all incoming patients: the VPP unit is suitable for patients who can be served "vertically" as opposed to "horizontally." Put differently, the VPP unit is only appropriate for patients who do not need an ED bed—the mode of service when a patient is on an ED bed is called "horizontal" since the patient is often laying down. Thus, the VPP design is a patient routing mechanism that relies on upfront predictions of who can be served without needing an ED bed. In contrast, PIT does not involve any prediction-based routing.

1.2. Organization and Summary of Contributions

Our contributions can be summarized as follows:

- We train an ML model on data from our partner hospital and validate it. Our ML achieves an out-of-sample Area Under the Receiver Operator Curve (AUC) of 88.7% and reveals that the most important patient features in predicting the need for an ED bed are ESI (negative association), age (positive association) and presence of abdominal chief complaints (positive association).
- We combine our ML model with a patient flow model to determine the optimal protocol for utilizing the VPP unit (based on both the characteristics of the ML model and those of the ED).
- We develop a realistic simulation and calibrate it with hospital data from our partner institution to examine the benefits of the VPP flow. We find that, if used under the optimal protocol and in conjecture with our ML model, the VPP design can lead to significant improvements in the average Length of Stay (LOS) and wait times compared to the existing practice. We also make use of our simulation to generate insights for other hospitals into whether and when they should adopt the VPP design.

[2] When this happens, the same evaluating physician often remains assigned to the patient so as to limit the rework, which is different from the FT and PIT approaches.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

- We also compare the VPP design with other ED flow designs such as FT and PIT, and find that the VPP design can outperform other flow designs, especially in hospitals where the proportion of low acuity patients is low. Our results indicate that in hospitals with a high prevalence of non-acute patients, a PIT design often achieves the best performance. We also find that, under some conditions, the FT design can lead to comparable performance to the VPP design. Finally, our analyses show that the VPP design is more robust than both FT and PIT designs due to its adaptive nature to changes in patient population characteristics, including age and ESI levels.

The remainder of the paper is structured as follows. In Section 2, we describe the study setting at our partner hospital, summarizing our data and the existing VPP implementation. We outline the related literature in Section 3. Section 4 presents a stylized model of the ED flow with a VPP unit. Section 5 characterizes the optimal policy for VPP usage. In Section 6, we develop and compare ML models that can determine VPP eligibility and illustrate the clinical insights that inform the VPP design. In Section 7, we make use of a simulation model calibrated with the data from our partner hospital to compare different patient prioritization rules within the optimal routing policy for the VPP unit. Section 8 compares the optimal VPP design with the FT and PIT flow models, identifying the relative merits of each approach. Finally, we conclude in Section 9 with a summary of our overall insights about whether, when, and how a VPP unit can help EDs improve their performance.

2. The VPP Unit at the Mayo Clinic

Prior to studying optimal ways of utilizing the VPP unit, we begin with a high level overview of the patient population at the Mayo Clinic Arizona ED and the current implementation of the existing VPP unit. Section 2.1 describes the operational characteristics of the ED while Section 2.2 focuses on the current implementation of the VPP unit.

2.1. The Emergency Department

The ED at the Mayo Clinic Arizona has 26 single treatment rooms, up to 9 hallway spaces, and is staffed with board-certified emergency physicians. To study its operations and ground our analysis to a real-world setting, we curated a retrospective dataset of routinely gathered ED operational data from the hospital's electronic health records database. Our dataset comprised de-identified records of all 49,350 patients who were served at the

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

Variable	Mean	Std
ESI	2.8	0.67
Age*	55.1	19.3
Visited VPP	6.3%	2.4%
Received ED bed	98%	13%
Received ED bed after VPP	72%	45%
Length of stay (mins)	238	119.4

Table 2 Summary of Mayo ED records.

Note: *Patients who are 85 or older are listed as 85+ in our data set for privacy protection.

ED between October 7th, 2018 and the 31st of December, 2019. This time period coincides with the initiation of a new electronic medical record and excludes visits seen during the coronavirus pandemic. The Mayo Clinic review board, along with our host academic institution, approved this protocol as minimal-risk research and waived the requirement for informed consent.

Our data shows that the average LOS in the ED is 238 minutes, with a standard deviation of 119 minutes (see Table 2). The peak demand hours for the ED were, on average, between 11:00 am and 2:00 pm, similar to the majority of other EDs (Lucero et al. 2021). We observe that 71.94% of the patients who are first seen at the VPP unit are subsequently assigned a bed in the ED, which shows a low accuracy in the current practice of identifying patients that can be served in the VPP without needing an ED bed (i.e., patients that can be served "vertically" and not "horizontally"). For those patients who were first seen in the VPP unit, the average (standard deviation) LOS in the ED is 3.82 (2.96) hours. Correspondingly, the average (standard deviation) LOS in the ED is 4.17 (2.83) hours for patients who were not routed to the VPP unit. Figure 2 illustrates the average arrival rate to the main ED as well as the fraction of patients that are routed to the VPP unit per hour of the day. As indicated in this figure, the VPP unit is open every day between 7:00 am and 11:00 pm. The exact opening and closing times may vary on a individual day basis depending on nursing staff availability.

2.2. Current Implementation of the VPP Unit

In the current practice, physicians identify potential VPP patients from their assigned patient load and contact a nurse who moves the patient into the VPP room for this assessment. The selection process for the VPP remains ad-hoc within the system and depends on the physician's perception regarding the patient's condition. Specifically, if the physician believes that a given patient in the waiting room can be served without the need

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

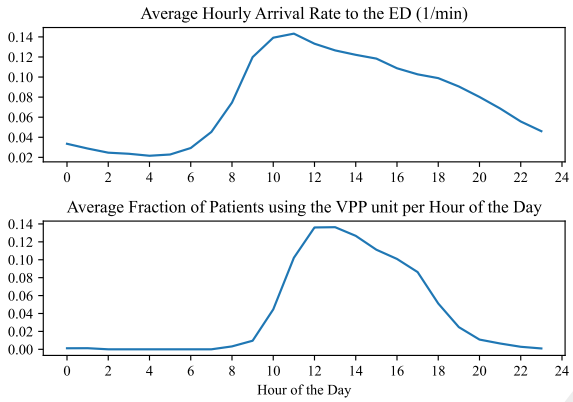


Figure 2 Average hourly arrival rate to the main ED and the VPP unit.

for intravenous medication or other types of treatment that requires an ED bed, she can request to see the patient in the VPP unit. Some physicians also use the VPP area not to fully treat and discharge a patient but to initiate a first set of results for those patients who they believe would benefit from an earlier assessment. After being seen in the VPP, patients return to the waiting room. However, if a bed becomes available, they are moved to that ED room or to a hallway bed. When all care can be facilitated from the VPP unit, patients get directly discharged from the VPP. The VPP unit—a single room with a single patient capacity—is located next to the waiting area in lieu of another ED ward in close proximity to all the other beds available in the department as shown in Figure 3.

The VPP unit provides a tool for the ED to quickly process patients that require minimal care in times of high traffic. However, when patients, that require more sophisticated care that can be provided only in a main ED bed, are originally routed through the VPP, further delays might be created in their treatment. In light of this, as the ED is moving to a new building, the hope is to obtain the ability to identify patients who can be fully served without needing an ED bed and route them the VPP unit. Currently, the decision of routing patients to the VPP unit lies on the shoulders of the ED physicians in a period and environment of high stress, and there are no specific protocols or set criteria for this purpose. Furthermore, the physicians have at their disposal only patient information at the time of triage, and thus, their decision-making remains a challenging task with a high frequency of errors. As discussed in Section 1, one of our main goals is to assist decision-making in this regard.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

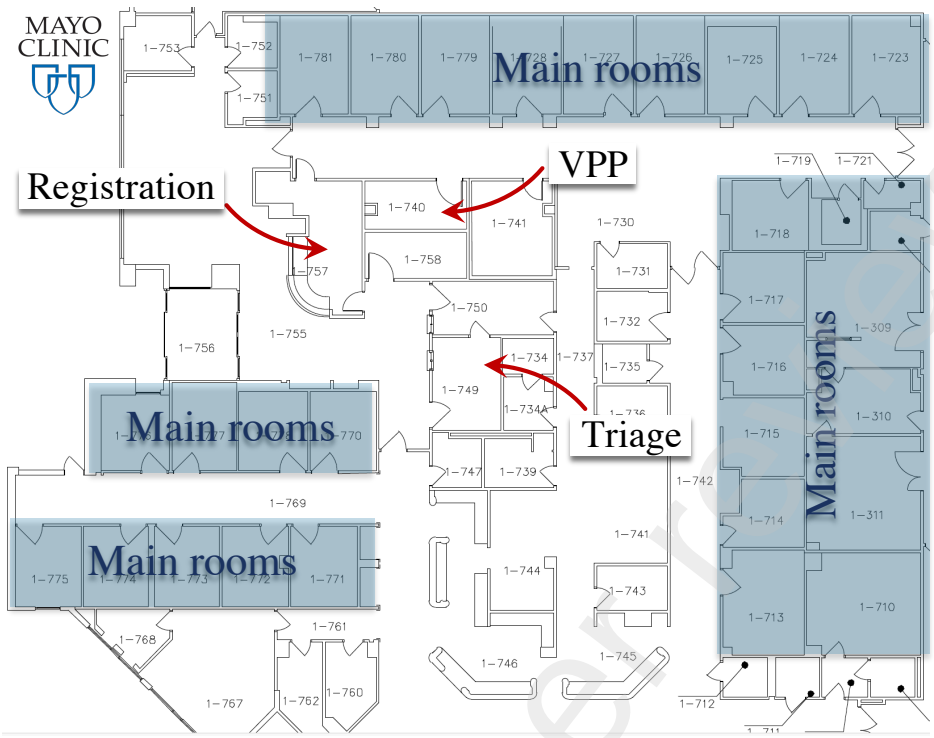


Figure 3 **Physical layout of the Mayo Clinic ED.**

3. Related Literature

There are two main streams of research related to this paper: (1) data-driven models to predict patient flow and outcomes in the ED, and (2) queueing models of patient routing within ED operations. We highlight the key contributions and findings from each of them below. For a complete review of operations research and management tools applied to ED patient flow, we refer interested readers to Saghafian et al. (2015), and the references therein.

Since the establishment of the Health Information Technology for Economic and Clinical Health Act (US Department of Health et al. 2014), hospitals have been consistently recording the trajectory of the patients within the ED in the form of electronic health records (EHRs). Such abundance of healthcare data has permitted the development of an increasing number of predictive models that attempt to estimate a patient's LOS in the ED. Chaou et al. (2017), Yoon et al. (2003) identified acuity level, age, and the need for additional tests (e.g., laboratory, X-rays, CT scans) as the most predictive factors for a longer LOS using multivariate logistic regression models. Gill et al. (2018) focused on FT patients in the Australian healthcare system to identify the reasons that lead to prolonged LOS. Similar to the previous work, the derived ML model showed that the most important

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

variables are the time taken from a patient's arrival to the time of ordering additional tests, potential admission to the hospital, and bed assignment. These findings indicate that the efficiency of ED streaming processes can be significantly affected when patients with more involved care needs are not appropriately routed. The importance of accurate triage on LOS was also highlighted by Partovi et al. (2001), who showed that a PIT model can offer a moderate decrease in ED LOS, although it is associated with relatively high costs. Nonetheless, there is a high degree of variation in LOS across ED physicians (Traub et al. 2018), making it a challenging outcome measure to predict at the time of triage accurately.

In addition to LOS, several studies have leveraged data-driven methodologies to predict other patient outcomes associated with ED visits at the time of triage. Hong et al. (2018) used data from three hospital systems to predict hospital admission at the time of ED triage, achieving an out-of-sample AUC of 87%. Raita et al. (2019) developed ML models with equivalent performance on predicting hospital admission and a slightly lower AUC (85%) on predicting intensive care unit (ICU) admission and mortality. Several other studies have been published on predicting mortality at the time of ED arrival either for the entire population or for specific diseases, using triage information (Lee et al. 2020, Bertsimas et al. 2020a, Klug et al. 2020). However, to our knowledge, there has been no study that proposes a validated approach to detect whether a patient visiting the ED will need care on a bed (i.e., "horizontal" treatment) or not (i.e., "vertical" treatment). One of the goals of our study is to address this gap in the literature.

From a queuing perspective, models for improving ED patient flow can be classified based on whether their goal is to reduce boarding times (i.e., delay from when an ED patient is admitted for inpatient care until they physically depart the ED) or to improve patient flow prior to being either admitted for inpatient care or discharged to go home. We refer readers to Feizi et al. (2022), Izady and Mohamed (2021), and (Saghafian et al. 2023) for reviews of the former but focus on the latter in this paper since the VPP has a similar objective.

Most related to our paper are Saghafian et al. (2012, 2014, 2018) and Li et al. (2021). Saghafian et al. (2012) uses a combination of a queueing model and simulations to determine when it is optimal to use a disposition-based patient streaming policy in the ED and the conditions under which this policy would result in maximum performance. Saghafian et al. (2014) proposes a complexity-augmented triage algorithm and demonstrates that

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

including an estimation of patient complexity in the traditional ED triage and patient streaming policies results in higher patient safety and lower overall LOS. Saghafian et al. (2018) studies optimal teletriage designs in which ED triage can be done by physicians in other physical locations and highlights the tradeoffs between speed and quality in making patient routing decisions. Following an observation that discharge patients are prioritized when the ED's blocking level (i.e., number of boarding patients) is high, Li et al. (2021) formulate a Markov Decision Process (MDP) and uses it to find a better patient prioritization policy based on patients' urgency and the blocking level information. Our paper adds to this literature by (a) investigating a VPP-based patient routing policy, which, to our knowledge, has not been studied analytically, and (b) showcasing how an ML model with a given level of accuracy can be implemented in an ED to improve its operational efficiency while considering its main system parameters (e.g., arrival and service rates).

4. An Analytical Model of the ED

This section aims to characterize the best protocol for using the VPP unit. Since all arriving patients at our partner hospital are randomly assigned to physicians through a rotational patient assignment algorithm, physicians have a fair and equal load of patients. Thus, to gain clear insights, we start by focusing on a single physician who balances her workload between the VPP and main ED. In Section 7, we further test the validity of our findings by developing and using a realistic simulation environment calibrated with hospital data.

4.1. Model Description

Patient Flow. The simplified patient-flow diagram for an ED physician serving arriving patients is depicted in Figure 4, which consists of two main components: the VPP unit and the main ED. Patients arrive with interarrival times drawn from an exponential distribution at a rate of $\lambda \in (0,1)$. For analytical tractability, we start by assuming that λ is time-homogeneous, but we relax this assumption in our data-driven simulation analyses (Section 7). A fraction of the arriving patients denoted by τ are sent to the VPP unit, and the rest, $(1-\tau)$, are routed to the main ED. The VPP unit and the main ED are assumed to operate with service rates denoted by μ_V and μ_E, respectively. Without loss of generality, we assume that μ_E is normalized to one ($\mu_E = 1$) and $\mu_V \geq 1$, which reflects the fact that the patients sent to VPP are served much faster than those seen in the main ED area.

A fraction, denoted by $p(\tau)$, of patients sent to the VPP unit will end up needing treatment in the main ED. Effectively, these patients endure the length of stay (LOS) of

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

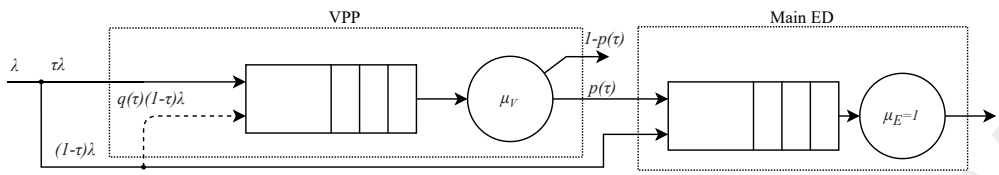


Figure 4 **Simplified model of the ED flow with two units: main ED and the VPP.**

both the VPP and main ED. Similarly, a fraction of patients denoted by $q(\tau)$ who were sent to the main ED could have been served in the VPP and discharged.[3] We note that $p(\cdot)$ and $q(\cdot)$ are functions of τ, because they depend on the type of patients initially routed to the VPP. We discuss how $p(\tau)$ and $q(\tau)$ depend on τ in Section 4.2 (see, e.g., Equations 7 and 8).

In practice, the patients in the main ED are prioritized and the physician will visit the VPP only when she has idle time. Specifically, apart from occasions in which the main ED queue is empty, idle time could also occur in practice when the physician is waiting for a patient's test results or the patient is administered a lengthy treatment such as an intervenous (IV) drug and requires no direct physician intervention until the treatment is over. In such occasions, although a main ED bed is occupied, the physician can attend to the VPP without compromising the LOS of the main ED patient. In contrast, VPP patients always require physician presence since the VPP is designed for physicians to conduct quick diagnosis and not lengthy treatments. Thus, we assume that the VPP operates as a queue with vacations (i.e., the server cannot serve waiting customers for periods of time), while the main ED operates as a queue without vacations.

In practical terms, this implies that the physician will serve her VPP patients, leave the VPP to perform other tasks (i.e., won't be able to serve VPP patients), and then return to the VPP once again. For tractability, we also start our analyses by assuming that all time durations (interarrival times, service times, and vacation times) are exponentially distributed. Under these assumptions, the VPP queue depicted in Figure 4 is an $M/M/1$ queue with exponential vacations. The average waiting time in such a queue, $\mathbb{E}[W_V]$, is calculated in Servi and Finn (2002), which in our setting translates to:

$$\mathbb{E}[W_V] = \frac{1}{\mu_V - \tau\lambda} + u, \tag{1}$$

[3] In our partner hospital, patients are not routed from the main ED to the VPP unit when there is a realization that the patient could have been served in the VPP unit. However, $q(\tau)$ represents an opportunity cost, which must be considered when deriving the optimal τ.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

where u is the average duration of the "vacation." Note that the average vacation length may be a function of how busy the main ED is; namely, if the physician is busy in the main ED she will visit the VPP less frequently. However, we assume u is exogenous since, in practice, it only depends on how frequently the physician decides to visit the VPP. In other words, although the main ED patients are prioritized, the priority rule is not such that the VPP is completely omitted during the times in which the main ED is extremely busy.

To find the waiting time in the main ED, we note that it has an arrival stream that is composed of two distributions: a Poisson process with a rate of $(1-\tau)\lambda$ and a non-Poisson process that is based on departures from the $M/M/1$ queue with vacation, as derived in Tang (1994). Using these two distributions, we derive the distribution of inter-arrival times to the main ED in Lemma 1.

Lemma 1. *Let $f_a(t)$ be the distribution of inter-arrival times to the main ED. We have:*

$$f_a(t) = -\frac{\lambda}{(1-\mu_V u)\Big[p(\tau)^2\lambda^2\tau^2u^2-1\Big]} \Bigg[(-1+\tau-p(\tau)\tau)(-1+\mu_V u)e^{(-1+\tau-p(\tau)\tau)\lambda t} + p(\tau)\tau(\mu_V-\lambda p(\tau)\tau)u(1+(1-\tau)\lambda u)e^{-\left((1-\tau)\lambda+1/u\right)t} + \mu_V p(\tau)\tau((1-\tau)\lambda+\mu_V)u^2(-1+\lambda p(\tau)\tau u)e^{-\left((1-\tau)\lambda+\mu_V\right)t}\Bigg]. \quad (2)$$

The proof of Lemma 1 and all the other proofs are presented in the Appendix. Note that Equations 1 and 2 denote the actual VPP wait time and main ED inter-arrival distribution, respectively, and not the counterfactual scenarios that would result from considering the $q(\tau)(1-\tau)$ portion of system arrivals that could have been served in the VPP.

We next use Kingman's approximation for the average wait time of a $G/G/1$ queue to estimate the main ED's average wait time, W_E. Specifically, we have:

$$\mathbb{E}[W_E] \approxeq \frac{\rho}{1-\rho}\Big(\frac{C_a^2+C_s^2}{2}\Big), \quad (3)$$

where $\rho = \tau\lambda/\mu_V$ is the utilization rate of the VPP, C_a^2 is the squared coefficient of variation (SCV) of inter-arrivals which can be calculated from Equation 2 ($C_a^2 = \mathrm{Var}[T_a]/\mathbb{E}[T_a]^2$,

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

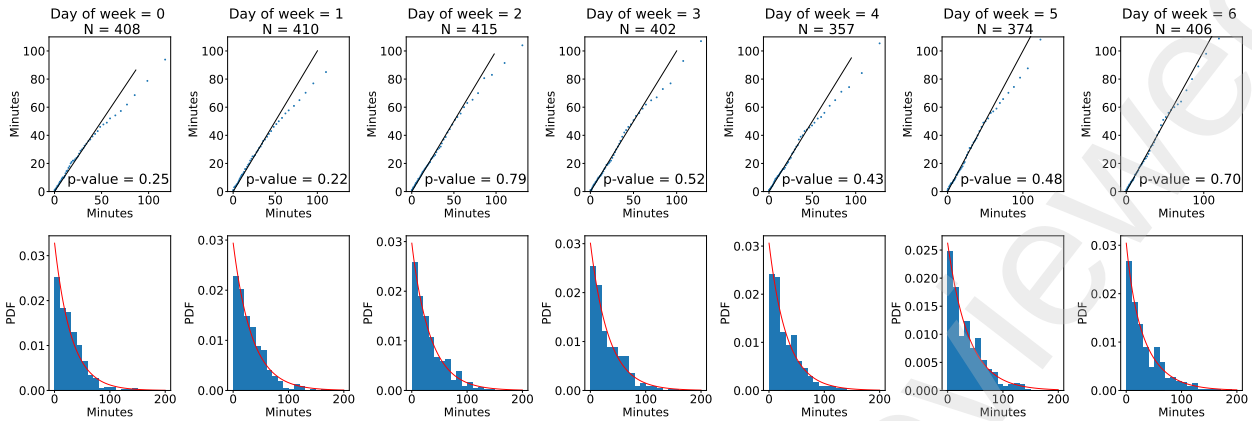


Figure 5 **Matching interarrival time distribution of main ED with Equation 2.**

where T_a denotes the inter-arrival times to the main ED), and $C_s^2 = 1$ is the SCV of service times. (Recall that the main ED's service time is exponentially distributed.) Again, similar to Equations 1 and 2, ρ is defined here for the actual system and not the hypothetical scenario in which the $q(\tau)(1-\tau)$ portion of the main ED patients are also served in the VPP.

Remark 1. We use the empirical arrivals to the main ED in our data to demonstrate the validity of the distribution derived in Equation 2. Specifically, we sort the actual arrivals to the main ED by physician, then date/time, and subsequently focus on the empirical distribution of arrivals for the time of day with peak arrival rates, 11 am - 12 pm, on each day of the week, separately[4]. As our data does not contain vacation lengths or the specific LOS in the VPP, we estimate these two parameters within reasonable practical ranges and show that Equation 2 matches what we observe from our data using the Kolmogorov–Smirnov (KS) test. The results are shown in Figure 5. The p-values from the KS-tests are denoted in the upper plots. They indicate that what we obtain from Equation 2 closely matches our empirical data.

Patients. We assume each arriving ED patient has a true and unknown medical complexity level that can be used for patient streaming (see, e.g., Saghafian et al. (2014)) denoted by $\gamma \in [0,1]$. A patient will need to be treated in the main ED, if their complexity level exceeds a threshold, α. In practice, α represents the baseline prevalence of patients who can be discharged directly from the VPP. At our partner hospital, $\alpha = 0.205$. However, to provide

[4] We focus on the busiest time of the day because the data density is higher and allows for a better approximation of the distribution.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

general insights useful for a range of EDs, in our model, we assume that $0 < \alpha < 0.5$, which implies that no more than half of the patients can be safely discharged from the VPP without needing a bed in the main ED. It is a reasonable assumption, because in any given ED it is extremely unlikely that over half of the patients will not require main ED care.

When patients with $\gamma > \alpha$ are routed to the VPP, they will need a bed in the main ED after their VPP service. Similarly, when patients with $\gamma \leq \alpha$ are routed to the main ED, they could have been discharged from the VPP. Based on these, we next develop a decision support tool that makes use of a machine learning model, and aids decision-making by determining the best routing policy considering the associated costs of misrouting patients.

4.2. A Decision Support Tool

Recall from Figure 4 that VPP patients need to be re-routed to the main ED with probability $p(\tau)$, and main ED patients could have been discharged from the VPP with probability $q(\tau)$. To derive $p(\tau)$ and $q(\tau)$, and hence the associated costs of misrouting patients, we first discuss a machine learning model that can be used as part of the decision support tool.

Machine Learning Model. We consider an ML that uses up-front patient features (e.g., triage information) to predict for each patient i a class label, $Y_i \in \{0, 1\}$, defined as whether they will need a bed in the main ED ($Y_i = 1$) or can be safely discharged from the VPP without eventually needing an ED bed ($Y_i = 0$).

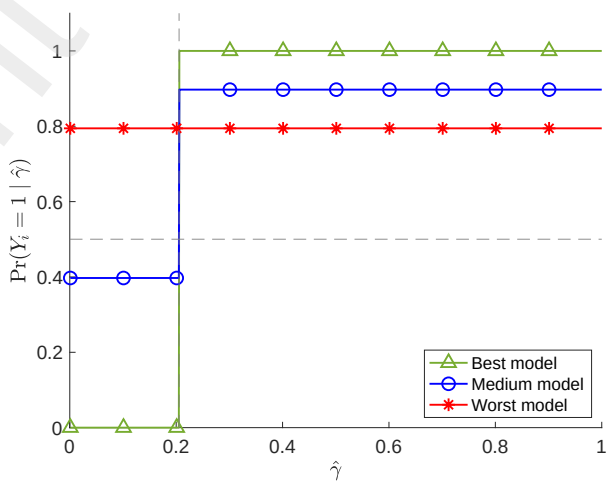


Figure 6 **Probability of needing an ED bed as a function of the estimated complexity score $\hat{\gamma}$.**

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

In essence, while the true complexity of the patient (γ) is unknown, the ML model maps the up-front patient information available into a predicted complexity, $\hat{\gamma}$. A threshold of τ is then used to route patients; patients with $\hat{\gamma} \leq \tau$ are sent to the VPP unit and the rest are routed to the main ED. A perfect ML model would predict $\hat{\gamma}$ such that a classification threshold of $\tau = \alpha$ would separate the two classes with 100% accuracy. Specifically, the model would suggest that $\Pr(Y_i = 1|\hat{\gamma})$ is equal to 1 for any $\hat{\gamma} > \alpha$, and is 0 otherwise. For an arbitrary ML model, we assume that the probability of needing a main ED bed is a piece-wise constant function of $\hat{\gamma}$ as stated in Equation 4 and depicted in Figure 6 for $\alpha = 0.205$[5]. Essentially, this models the behavior of any ML model with a non-linear shape (e.g., Sigmoid), in which the highest F1-score is achieved when the classification threshold is set to α.[6]

$$\Pr(Y_i = 1 \mid \hat{\gamma}) = \begin{cases} k_1, & \text{if } \hat{\gamma} \leq \alpha \\ k_2, & \text{if } \hat{\gamma} > \alpha \end{cases} \tag{4}$$

In Equation 4, $k_1 \in (0, 1-\alpha)$ is a constant that describes the model's quality. k_2 is calculated based on the fact that, ultimately, $1-\alpha$ patients need a main ED bed, regardless of what prediction model is used. Namely:

$$\int_0^1 \Pr(Y_i = 1 \mid \hat{\gamma}) \mathrm{d}\hat{\gamma} = 1 - \alpha \ \therefore \ k_2 = 1 - \frac{k_1 \alpha}{1-\alpha}. \tag{5}$$

It is helpful to discuss the properties of the assumed ML model in Equation 4 for further clarification. First, $k_1 \to 0$ represents a "perfect model" with an Area Under Curve (AUC) of $AUC \to 1$. Conversely, if the model makes predictions randomly (i.e., "worst model") we have $k_1 \to 1-\alpha$ and $AUC \to 0.5$. In the latter case, $\hat{\gamma}$ has no meaningful relationship with γ so the probability of needing a bed in the main ED is equal to the baseline probability for all $\hat{\gamma}$ values. Second, the AUC of the model, in general, can be calculated using Lemma 2.

Lemma 2. *For any $\alpha \in (0,1)$ and $k_1 \in (0, 1-\alpha)$, we have:*

$$AUC = 1 - \frac{k_1}{2(1-\alpha)}. \tag{6}$$

[5] As mentioned earlier, 20.5% of patients are seen in the VPP unit at our partner hospital. Thus, we set $\alpha = 0.205$ in Figure 6.

[6] Recall that α is a threshold level of patient complexity, below which patients do not need a main ED bed. Thus, an effective ML model will separate patients with $\gamma > \alpha$ from those with $\gamma \leq \alpha$.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

Remark 2. While we use k_1 and α as the main ML model parameters, we present the results in Section 5 using an AUC measure for higher clarity. Furthermore, without loss of generality, we may assume that $\hat{\gamma}$ comes from a uniform distribution. To implement in practice, $\hat{\gamma}$ can be thought of as quantiles, deciles, or any other equal division of the predicted values. For example, if $\tau = 0.15$, this is equivalent to routing the bottom 15% of $\hat{\gamma}$ values to the VPP regardless of the true underlying distribution of $\hat{\gamma}$.

Probabilities of Misrouting ($p(\tau)$ and $q(\tau)$). We state that a patient is *misrouted* if they are sent to the VPP but then needed a main ED bed, or if they were sent to the main ED but could have been seen and discharged from the VPP (i.e., without needing an ED bed). With $\Pr(Y_i = 1 \mid \hat{\gamma})$ defined, we can now derive the system-level conditional probability that a patient needs a main ED bed given that she has been routed to the VPP. For a given threshold, τ, $p(\tau)$ is derived in Equation 7.

$$p(\tau) = \Pr(Y_i = 1 \mid \text{VPP}, \tau) = \frac{1}{\tau}\int_0^{\tau} \Pr(Y_i = 1 \mid \hat{\gamma}) d\hat{\gamma} = \begin{cases} k_1, & \text{if } \tau \leq \alpha \\ \dfrac{(\tau-\alpha)(1-\alpha) + \alpha k_1 (1-\tau)}{(1-\alpha)\tau}, & \text{if } \tau > \alpha \end{cases} \tag{7}$$

Similarly, the system-level probability that a patient sent to the main ED could have been served in the VPP is given by $q(\tau)$ is derived in Equation 8.

$$q(\tau) = \Pr(Y_i = 0 \mid \text{Main ED}, \tau) = \frac{1}{1-\tau}\int_{\tau}^{1} \Pr(Y_i = 0 \mid \hat{\gamma}) d\hat{\gamma} = \begin{cases} \dfrac{\alpha - \tau(1-k_1)}{1-\tau}, & \text{if } \tau \leq \alpha \\ \dfrac{\alpha k_1}{1-\alpha}, & \text{if } \tau > \alpha \end{cases} \tag{8}$$

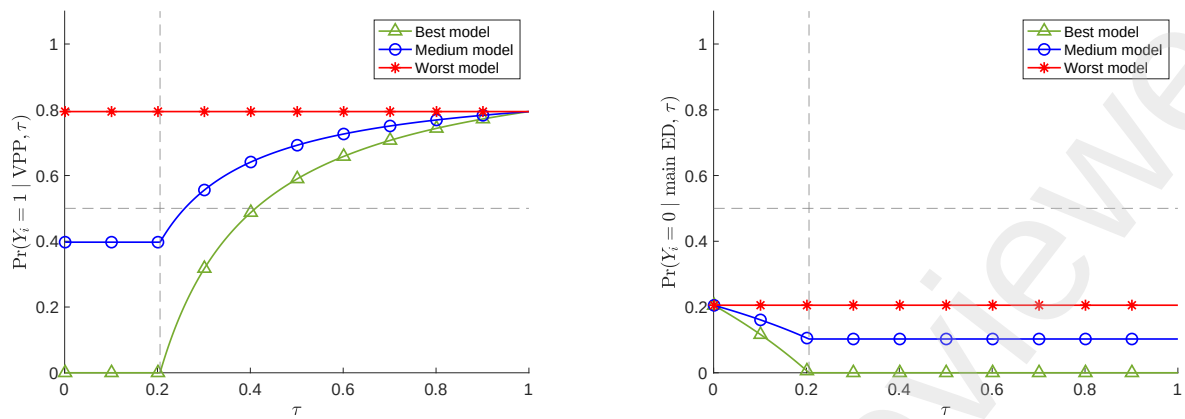
Figures 7a and 7b depict the $p(\cdot)$ and $q(\cdot)$ functions, respectively, for different values of k_1 while setting $\alpha = 0.205$. Note that with the best model, $k_1 \approx 0$, $p(\alpha) \approx q(\alpha) \approx 0$, as expected; and in the worst model equivalent to random selection, $k_1 = 1-\alpha$, $p(\tau) = 1-\alpha$ and $q(\tau) = \alpha$ which denote the baseline probabilities of needing and not needing a main ED bed, respectively.

4.3. Cost Function

To generate insight into effective routing protocols, we consider a model in which the decision-maker's goal is to minimize the associated costs of misrouting patients with the

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161



(a) Probability of needing an ED bed, given that a patient is served in the VPP.

(b) Probability of not needing a bed, given that a patient is served in the main ED.

Figure 7 Illustration of the $p(\cdot)$ and $q(\cdot)$ functions.

overarching goal of reducing the overall LOS in the ED. The total cost due to patient misrouting is the sum of the additional LOS incurred in the system due to type I and II misrouting errors, which for clarify we refer to as the over-utilization and under-utilization of the VPP.

The over-utilization cost is associated with patients who need a main ED bed but are misrouted to the VPP. Since they will be sent to main ED after being served in the VPP unit, they experience both the LOS of the VPP and main ED. However, they could, ideally, experience only the LOS of the main ED. Hence, the cost associated with over-utilization of the VPP is:

$$C_O(\tau \mid \alpha, k_1, \lambda, \mu_V) = \underbrace{\left(L_V \Big|_{\lambda_V = \tau\lambda} + L_E \Big|_{\lambda_E = (1-\tau)\lambda + p(\tau)\tau\lambda} \right)}_{\text{need main ED bed}} - \underbrace{\left(L_E \Big|_{\lambda_E = (1-\tau)\lambda + p(\tau)\tau\lambda} \right)}_{\text{if routed to main ED}} = L_V \Big|_{\lambda_V = \tau\lambda}, \tag{9}$$

where C_O denotes the over-utilization cost, and L_E and L_V represent the LOS of the ED and VPP, respectively. Important to note is that the arrival rate of the ED would not change even if the patients were routed correctly. This is because the patient flow is such that all patients who need treatment in the main ED are eventually routed there. As such, the over-use cost is simplified to the LOS of the VPP.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

The under-utilization cost relates to patients who were routed to the main ED while they could have been entirely served in the VPP. For these cases, the cost function is the difference of their expected LOS in the main ED with that of the VPP. Thus, we have:

$$C_U(\tau \mid \alpha, k_1, \lambda, \mu_V) = \underbrace{L_E\Big|_{\lambda_E=(1-\tau)\lambda+p(\tau)\tau\lambda}}_{\text{could be seen in VPP}} - \underbrace{L_V\Big|_{\lambda_V=\tau\lambda+q(\tau)(1-\tau)\lambda}}_{\text{if routed to VPP}}, \tag{10}$$

where $q(\tau)(1-\tau)\lambda$ in λ_V is added because the hypothetical scenario must consider the additional load to the VPP.

For a given machine learning model and set of system parameters, our goal is to find the optimal threshold τ^* for routing patients to the VPP. Note that $p(\tau)\tau$ and $q(\tau)(1-\tau)$ portion of the patients experience the over-use and under-use costs, respectively. Hence, τ^* minimizes the overall cost function and can be written as:

$$\tau^* = \arg\min_{\tau\in[0,1]}\{p(\tau)\tau C_O(\tau \mid \alpha, k_1, \lambda, \mu_V) + q(\tau)(1-\tau)C_U(\tau \mid \alpha, k_1, \lambda, \mu_V)\}. \tag{11}$$

5. Model Results

In this section, we derive the optimal threshold τ^*, and discuss its main properties. Combining Equations 2 and 3 to gain meaningful insights is mathematically intractable in general. Therefore, in what follows, we first generate insights by assuming that $u=0$, but relax this in our simulation analyses in Section 7.

Solving Equation 11, we can find a unique solution for any given ML model and set of system parameters as stated in the following result.

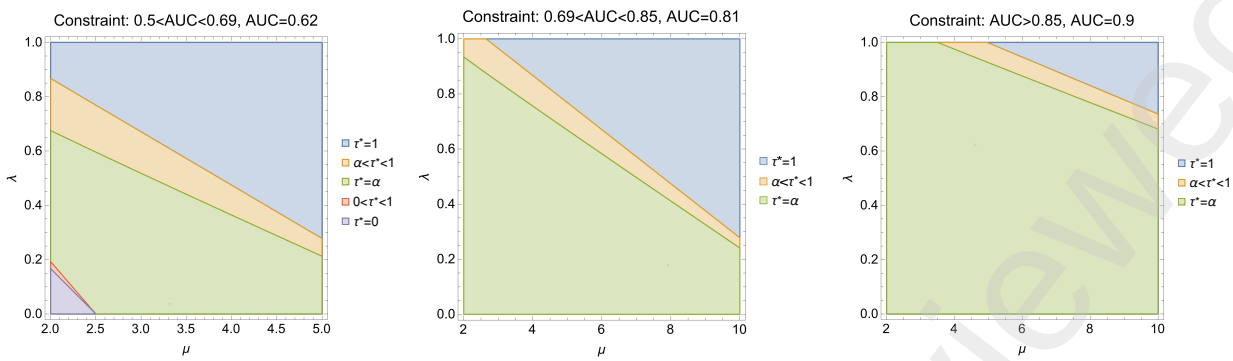
THEOREM 1. *For any combination of* $(\mu_V, \lambda, \alpha, k_1)$, τ^* *defined in Equation 11 is unique and is given in Table EC.1.*

Figures 8 shows the optimal threshold, τ^*, as a function of the arrival rate and VPP service rate, μ_V, for different levels of AUC. Theorem 1 enables us to make several key observations.

PROPOSITION 1. *In the* (μ_V, λ) *space, denote the region in which* $\tau^* = \alpha$ *by* $\mathcal{A}_k$ *for* $k_1 = k$. *If* $k < k'$, *then* $\mathcal{A}_{k'} \subset \mathcal{A}_k$.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161



(a) Low-AUC (b) Medium-AUC (c) High-AUC

Figure 8 Region plots showing τ^* as functions of λ and μ_V for $\alpha = 0.206$.

Recall that a lower k_1 value corresponds to a higher AUC (Equation 6). Proposition 1 states that making use of a better ML model (with higher AUC) reduces the need for "risk-taking" for larger combinations of (λ, μ_V) because a higher accuracy in the model essentially means that there is less filtering required at the VPP, and less opportunity cost at the main ED. This implies that the classification threshold that maximizes the F1-score of the ML model is also the optimal operational patient routing threshold for a larger combination of (λ, μ_V) values.

Lemma 3 below describes the relationship between τ^* and the (λ, μ_V) values.

LEMMA 3. $\dfrac{\partial \tau^*}{\partial \mu_V} \geq 0$ *and* $\dfrac{\partial \tau^*}{\partial \lambda} \geq 0$.

Lemma 3 implies that as the VPP's service rate increases, it becomes optimal to use the VPP for more patients (all else equal). This is because the added LOS due to a (perhaps unnecessary) VPP visit is substantially less than the LOS of the main ED, and thus, the filtering mechanism of VPP in reducing the number of patients sent to the main ED becomes particularly useful. In addition, Lemma 3 states that as the overall arrival rate increases, more patients should be sent to the VPP unit (all else equal). Intuitively, this implies that in hospitals where the arrival rate is high, it is beneficial to overuse the VPP to filter some of the patients who can be discharged quickly—albeit with less certainty—in an attempt to reduce overcrowding in the main ED.

We also note from Figure 8 that for some combinations of parameters $\tau^* = 1$, meaning that all arriving patients should be first routed to the VPP. This implies conditions under which VPP should be used similar to the PIT (see Section 1), and occurs when the ML model's accuracy is below a threshold. Proposition 2 formalizes such conditions.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

PROPOSITION 2. *$\tau^* = 1$ if:*

- *$\mu_V > \mu_4(k_1, \alpha)$ for all λ and k_1 or;*
- *$\lambda > \lambda_4(k_1, \alpha, \mu_V)$ for $\mu_V > \max(2, \mu_2(k_1, \alpha))$*

where $\mu_2(\cdot)$, $\mu_4(\cdot)$, and $\lambda_4(\cdot)$ are defined as follows:

$$\mu_2(k_1, \alpha) = \frac{(1-\alpha)(1-k_1)^2}{k_1}$$
$$\mu_4(k_1, \alpha) = \frac{1-\alpha}{\alpha k_1}$$
$$\lambda_4(k_1, \alpha, \mu_V) = \frac{1-\alpha-\alpha k_1 \mu_V}{(1-\alpha)^2}$$

Proposition 2 describes the conditions under which the VPP unit should be used similar to a PIT model. This occurs when one of the following conditions holds: (1) The service rate of the VPP unit is faster than a threshold, $\mu_4(k_1, \alpha)$, which would justify the incurred LOS penalty of erroneously being served in the VPP with the hope of not having to incur the much longer main ED LOS. Since μ_4 is decreasing in k_1, this implies that as the predictive power of the ML model improves, the VPP must be even faster for it to be optimally used as a PIT. (2) The predictive power of the ML model is better than $k_A(\alpha)$ and the arrival rate is higher than $\lambda_4(k_1, \alpha, \mu_V)$; or (3) The predictive power of the ML model is worse than $k_A(\alpha)$, the arrival rate is above $\lambda_4(k_1, \alpha, \mu_V)$, and the VPP is faster than $\mu_2(k_1, \alpha)$. The second condition implies that if the ED's arrival rate increases beyond a threshold, the wait time of the main ED becomes so long that it is optimal to serve all patients in the VPP first with the hope of removing all of those who can be discharged directly via the VPP from the main ED queue. The third condition further implies that when the predictive power of the ML model is worse than $k_A(\alpha)$, the service rate of the VPP must be faster than $\mu_2(k_1, \alpha)$ for the overuse cost to be justified.

The following lemma establishes further insights into the structure of the optimal routing policy.

LEMMA 4. *$\dfrac{\partial \mu_2}{\partial k_1} < 0$, $\dfrac{\partial \mu_4}{\partial k_1} < 0$, and $\dfrac{\partial \lambda_4}{\partial k_1} < 0$.*

Lemma 4 states that as the predictive power of the ML model drops (i.e., as k_1 increases), the ED is better off using a PIT model under a wider set of system parameters. In other words, the VPP is useful compared to a PIT model only when it is used in conjunction with an ML model that has decent predictive power. We further test this finding in Section 7 using realistic simulation analyses calibrated with hospital data.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

Finally, the following proposition establishes conditions under which the VPP unit should not be used (i.e., $\tau^* = 0$).

PROPOSITION 3. *$\tau^* = 0$ iff $\mu_V < \mu_1(k_1, \alpha)$, $\lambda < \lambda_1(k_1, \alpha, \mu_V)$ and $1/2 < k_1 < 1 - \alpha$, where $\mu_1(\cdot)$ and $\lambda_1(\cdot)$ are defined as follows:*

$$\mu_1(k_1, \alpha) = \frac{1}{1 - k_1};$$
$$\lambda_1(k_1, \alpha, \mu_V) = \frac{1}{2}\left(2 - (1-\alpha)(1-k_1)\mu_V - \sqrt{(1-k_1)\mu_V\left(4\alpha + (1-\alpha)^2(1-k_1)\mu_V\right)}\right).$$

Proposition 3 suggests the following. When a weak ML model is used, patients routed to the VPP are highly misclassified and hence, need to be sent to the main ED after the VPP visit.[7] In addition, with a combination of low arrival rates and a slow VPP, it is not justifiable to risk routing patients to the VPP, especially when the main ED's LOS is sufficiently low.

In closing this section, we note that Equation 6 and Figure 8 show that for a variety of practical settings $\tau^* \in (0, 1)$, meaning that only a proportion of patients should be routed to the VPP unit. In what follows, we make use of hospital data and train ML models to gain further insights into the characteristics of such patients.

6. Predicting VPP Eligibility Using Machine Learning

Leveraging data from our partner hospital, we train and validate ML models that predict, for each arriving patient, whether they can be discharged by being served in the VPP unit only (VPP eligibility). In this section, we describe the curated dataset, the proposed ML models, and the clinical insights that we obtained from our analysis.

6.1. Data Description

Following standardized protocols across the country, the Mayo Clinic records patient demographic information, vital values, and a chief complaint from each patient. The decision of whether a patient can be examined in the VPP unit takes place after triage while the patient is in the waiting room. Thus, for the development and validation of the downstream ML models, we leverage only the limited information provided at the time of triage as potential predictors. Table 3 provides an overview of the clinical characteristics of the derived patient population, reporting the summary statistics for both continuous and

[7] Recall that under the worst model $p(\tau) = \alpha$.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

binary variables as well as the percentage of missing values. At the time of data extraction, chief complaints were grouped into 28 categories by the clinical team (see Table 3). The clustering was conducted purely based on the clinical relevance and did not involve any statistical methodology.

In addition to the clinical information at the time of triage, we received for each patient whether and what type of additional examinations were subsequently performed. We also obtained operational and discharge information regarding the timing and the part of the ED where care was provided. This allowed us to analyze the entire patient trajectory in the ED, and correspondingly calibrate our simulation models in Section 7.

6.2. Outcome of Interest

The outcome of interest for the target supervised ML model is whether a patient will require a bed at the main ED prior to discharge. From our data, however, we do not observe this variable for patients that *could have been* discharged without needing an ED bed. That is, we only observe in our data patients who are discharged after being seen in the VPP unit without being assigned an ED bed. However, a percentage of patients who are assigned an ED bed, could have been discharged without it, if they had been initially routed to the VPP unit (instead of the main ED). For this reason, and based on conversations with ED physicians, we identify patients that can be discharged without an ED bed using the following criteria:

- Patients who received care at the VPP and were subsequently discharged without the use of an ED bed ($N = 901$).
- Patients with ESI scores of 2 or 3 who were treated using an ED bed but were discharged home after their ED visit without being admitted to the hospital. In addition, we require for this population the provision that no intravenous medication or fluids were administered, neither X-rays, CT scans nor ultrasounds were performed during the ED visit. Moreover, we require that the time from first contact to discharge was at most 2 hours ($N = 4,382$).
- Patients with ESI scores of 4 or 5 who were treated using an ED bed, but were discharged home after their ED visit without being admitted to the hospital. Moreover, we require for this population the provision that no intravenous medication or fluids were administered during the hospital stay ($N = 4,861$).

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

Independent Variable	Type	Distribution Information	% Missing
Demographic Information			
Arrival Age	Numeric	61.0 (43.0-74.0)	0.00%
Race White	Binary	43672.0 (88.5%)	0.00%
Race Asian	Binary	1407.0 (2.9%)	0.00%
Race Black or African American	Binary	2037.0 (4.1%)	0.00%
Race Choose Not to Disclose	Binary	518.0 (1.0%)	0.00%
Race Other	Binary	1687.0 (3.4%)	0.00%
Gender Male	Binary	22950.0 (46.5%)	0.00%
Acuity Score and Vitals at Triage			
ESI	Numeric	3.0 (2.0-3.0)	0.10%
SPO2	Numeric	98.0 (96.0-99.0)	0.30%
Diastolic Blood Pressure at Triage	Numeric	80.0 (72.0-89.0)	0.60%
Pulse Rate at Triage	Numeric	83.0 (72.0-96.0)	0.50%
Respiratory Rate at Triage	Numeric	18.0 (16.0-20.0)	0.50%
Systolic Blood Pressure at Triage	Numeric	136.0 (121.0-153.0)	0.60%
Temperature at Triage	Numeric	36.7 (36.5-36.9)	2.40%
Chief Complaint Categories			
Abdominal Complaints	Binary	6456.0 (13.1%)	0.00%
Abnormal Test Results	Binary	1829.0 (3.7%)	0.00%
Allergic Reaction	Binary	262.0 (0.5%)	0.00%
Back or Flank Pain	Binary	2642.0 (5.4%)	0.00%
Breast Complaints	Binary	61.0 (0.1%)	0.00%
Cardiac Arrhythmias	Binary	1055.0 (2.1%)	0.00%
Chest Pain	Binary	3679.0 (7.5%)	0.00%
Dizziness/Lightheadedness/Syncope	Binary	1969.0 (4.0%)	0.00%
Ear Complaints	Binary	254.0 (0.5%)	0.00%
Epistaxis	Binary	260.0 (0.5%)	0.00%
Exposures, Bites, and Envenomations	Binary	261.0 (0.5%)	0.00%
Extremity Complaints	Binary	5389.0 (10.9%)	0.00%
Eye Complaints	Binary	730.0 (1.5%)	0.00%
Falls, Assaults, and Trauma	Binary	2399.0 (4.9%)	0.00%
Fatigue and Weakness	Binary	1548.0 (3.1%)	0.00%
Fevers, Sweats or Chills	Binary	1908.0 (3.9%)	0.00%
Gastrointestinal Issues	Binary	3359.0 (6.8%)	0.00%
Genital Complaints	Binary	683.0 (1.4%)	0.00%
Medical Device or Treatment Issue	Binary	481.0 (1.0%)	0.00%
Medication Request	Binary	76.0 (0.2%)	0.00%
Neurological Issue	Binary	3457.0 (7.0%)	0.00%
Other	Binary	808.0 (1.6%)	0.00%
Other Pain	Binary	794.0 (1.6%)	0.00%
Psychiatric Complaints	Binary	206.0 (0.4%)	0.00%
Shortness of Breath	Binary	3050.0 (6.2%)	0.00%
Skin Complaints	Binary	2347.0 (4.8%)	0.00%
Upper Respiratory Symptoms	Binary	1941.0 (3.9%)	0.00%
Urinary Complaints	Binary	1446.0 (2.9%)	0.00%

Table 3 Summary statistics of all patient characteristics for the total sample. For continuous variables, we report the average and the 95% confidence interval. In the case of binary variables, the table shows the count of observations where the feature is present and in parentheses the percent over the entire population. The last column includes the percent of missing values in the dataset for each independent variable.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

The above criteria have been developed and validated by the emergency physicians at our partner hospital, and are based on clinical insights regarding patients who can be served without being assigned an ED bed.

6.3. ML Models

Leveraging the data from our partner hospital, we train binary classification models to predict whether a patient's care will require the use of an ED bed. We compare a wide range of well-established ML algorithms, including logistic regression with regularization (to avoid overfitting), classification trees (CART), random forests, gradient boosted trees (XGBoost), support vector machines (SVM), and multi-layer perceptron (MLP) (Hastie et al. 2009, Breiman et al. 2017, Breiman 2001, Chen and Guestrin 2016, Cortes and Vapnik 1995, Rosenblatt 1958). To train these, we split the patient population into a training (75%) and a testing cohort (25%) using the Sciki-learn library Pedregosa et al. (2011). We ensure that the ratio of prevalence for the outcome does not vary between the two sets. We tune seven model parameters by maximizing the K-fold cross-validation AUC using a bayesian optimization framework Head et al. (2020). In Table 4, we report the average value and standard deviation of the AUC on the testing set for five independent partitions of the data.

Table 4 shows that overall all algorithms achieve an average AUC above 85%. Furthermore, we observe that the performance across the different algorithms is fairly stable. Specifically, the best performing algorithm (XGBoost, AUC=88.7%) differs only by 2.4 percentage points in mean AUC compared to the algorithm with the worst performance (CART, AUC=86.3%). We also observe that there is not much variability in terms of the reported AUC across different splits of the data, as indicated by the standard deviation metric.

Algorithm	Mean AUC	Std. of AUC
CART	0.86286	0.00951
Logistic Regression	0.86333	0.01033
Random Forests	0.87833	0.00983
XGBoost	0.88714	0.00756
MLP	0.88432	0.00781
SVM	0.86754	0.00432

Table 4 Mean and standard deviation of the AUC metric on the testing set across all ML algorithms considered. The reported numbers correspond to the average performance on five independent splits of the data for the binary outcome of interest.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

6.4. Clinical Insights

We employ the SHapley Additive exPlanations (SHAP) to identify the risk drivers associated with our outcome of interest Lundberg and Lee (2017), Lundberg et al. (2020). In Figure 9, we report the average SHAP value of the 20 most important feature predictors in our best ML model—XGBoost. These values are ordered by decreasing significance. Higher feature values are indicated in red and lower feature values are in blue. For binary features, a discrete scale is used, whereas for non-categorical features there is continuity on the color scale. The overall patient risk corresponds to the sum of the SHAP values of all the features. Positive SHAP values are positively correlated with higher chance of needing an ED bed and negative SHAP values indicate decreases in the probability of needing a bed. For example, in Figure 9, lower values of ESI (blue) yield higher SHAP values, suggesting that more acute cases are more likely to need an ED bed. In contrast, the SHAP value increases with higher values (red) of age, suggesting that higher values of age are associated with higher probability of requiring an ED bed.

It can be seen from Figure 9 that ESI is the most significant variable followed by age. Our analysis also indicates that patients that had a chief complaint involving abdominal pain, chest pain, shortness of breath, neurological issues, fatigue, weakness, fever, or falls and traumas are more likely to need an ED bed. In contrast, patients with skin, urinary, abnormal test results, eye, and extremity issues as their chief complaints are less likely to require an ED bed. In terms of other triage information, low diastolic, systolic blood pressure values and oxygen saturation increase the chance of needing an ED bed. The opposite trend applies to the cases of low triage pulse rate, respiratory rate, and body temperature.

Our findings are in line with other studies who focused on identifying critically ill patients that require significant care using triage information at the ED. Specifically, Hong et al. (2018), Sun et al. (2011) also highlight ESI score and age as two of the most predictive factors. Sun et al. (2011) also relates irregularities in blood pressure values and history of hypertension to patients at higher risk. Similar findings are uncovered by Zhang et al. (2017) with the use of natural language processing and neural networks. Raita et al. (2019) showcase the significance of vitals measurements, identifying a high association between oxygen saturation, respiratory rate, pulse rate, and systolic blood pressure and patient outcomes at the ED. Our models differ from the aforementioned studies in that they mainly

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

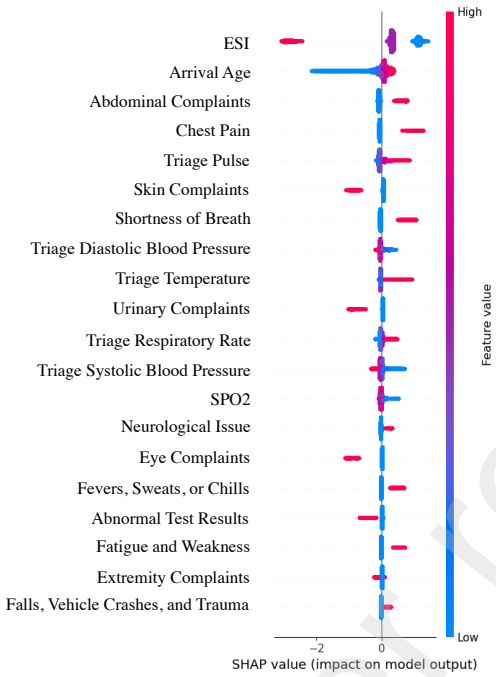


Figure 9 SHAP Plot for XGBoost models summarizing the contribution to risk prediction of the 20 most important features.

focus on predicting hospital admissions post-ED service while our focus is on predicting the need for an ED bed. In addition, contrary to some of these studies, our analysis did not involve any natural language processing, since chief complaints were clustered in broader, clinically relevant categories by the medical team.

7. Simulation of the ED Flow

To test the validity of the findings obtained from our simplified model (Section 5) and also to gain deeper insights, we developed a simulation model of the ED flow and calibrated it with hospital data. Section 7.1 describes the simulation model and steps taken in validating it. Section 7.2 showcases how to map the analytical model for the VPP presented in Sections 4-5 to a real-world ED and identify the resulting optimal routing policy. Section 7.3 highlights the differences between three distinct routing rules for the VPP. In Section 8, we then extend our simulation environment to (a) compare the VPP design with other ED flow approaches introduced in the Introduction (e.g., LOS and PIT), and (b) generate insights into when, and for which hospitals, the VPP design is advantageous.

7.1. Data-Driven Simulation Model: Development and Validation

We develop a simulation model of the Mayo Clinic ED based on the operational constraints of the system and the clinical characteristics of the population it serves. Our aim is to

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

design a realistic virtual test bed of the ED, where we can test the impact of different routing and prioritization protocols on patients' average LOS and waiting time.

Arrival Process. We assume that the arrivals to the ED follow a non-stationary Poisson process with a dynamically changing rate during the day, following the empirical arrival rates of the ED presented in Figure 2. The arrival rate in our simulations is specified for each ESI class, ranging from one to five, and for each hour of the day.

Patient Population. Each simulated arrival is sampled from a synthetic pool of patients based on the ED records, using the synthetic data vault (SDV) framework (Patki et al. 2016). The SDV process involves three consecutive parts: (1) data extraction and processing (DataNavigator); (2) generative model development (Modeler); (3) synthetic data creation (Sampler). SDV first estimates the distribution of each individual feature and the covariance between all independent variables in the dataset. Subsequently, the algorithm selects between a set of common distributions, including the truncated Gaussian, the uniform, or the beta distribution, the one that best matches the real data according to the p-values of the Kolmogorov-Smirnov test. Thus, the shape of the chosen cumulative distribution function for each feature is determined by the significance level of the statistical test. Finally, a Gaussian Copula function is applied to characterize the joint distribution of all derived random variables, ensuring that the shape of different distributions does not influence the covariance estimates. The SDV approach allows us to generate a realistic synthetic patient population that approximates the patient volume and mix that is served for each hour of the day at our partner hospital.

Simulating Assignments to VPP and ED (Current Practice). In the current practice, all arriving patients are assigned to a physician using a randomized round-robin algorithm. Patients are also triaged and then sent to the waiting area. By default, patients waiting will be taken to an ED bed and served by their assigned physician. However, when a physician becomes available, she considers the pool of patients assigned to her who are still in the waiting area and assesses whether they can be served in the VPP unit. If the physician decides that a patient can be served in the VPP, the physician requests that the patient be moved to the VPP unit. We model this ad-hoc selection as a Bernoulli process, where the probability of success (i.e., selection to the VPP) is a function of each patient's ESI level and hour of the day. We observe that this Bernoulli process matches our data relatively well (see Table EC.2). It also ensures that patients are served in the VPP (in the

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

simulated environment) only during the hours in which the VPP is open. Upon completion of the VPP visit, depending on the value of the test results, patients may either (a) be sent to the main ED queue for additional ED care, or (b) get discharged to go home directly from the VPP unit. The overall patient flow is based on Figure 1c.

Our simulation analyses extend the analytical framework presented earlier to consider a system that involves multiple physicians. We leverage the overall patient arrival rate to the ED and the average number of physicians working at any given hour from our data. Our approach considers the "competition" among physicians for utilizing the VPP, rendering it a shared resource in the ED.

Service Process. Once a patient has been seen in the VPP unit, tests are ordered. We assume that VPP patients will have to wait for their tests to be completed to determine whether they need to be served in the main ED. We extract disposition times and test times from our data and observe that for about half of the ED visits (49.5% for Main ED patients and 53.7% for VPP patients), a test has been ordered prior to the physician's first contact (see Fig. EC.1). Hence, we assume that a patient's service time begins from the earlier of first physician contact and first ordered test, and ends when a patient is either admitted to the hospital or discharged to go home. Also, we observe that for about 10% of patient visits, a test result becomes ready after the disposition decision is made. Since these have a low percentage of occurrence, we exclude them from the service duration and do not assume they cause further delays once the disposition decision has been made. Since ED service includes both treatment and testing, we also incorporate, as system parameters, the average testing and treatment durations separately for each ESI level and hour of the day. We assume that these durations follow time-varying exponential distributions, with means extracted from our data.

In addition, we model the probability that a patient is admitted to an inpatient unit after ED service based on ESI levels, and calibrate it using our data. Patients admitted to an inpatient unit after ED service often experience a "boarding time," which involves waiting in the ED until an inpatient bed becomes available. We model this using a log-normal distribution (see, e.g., Saghafian et al. (2023)) with means and standard deviations as functions of ESI and the hour of the day (obtained from our data).

Validation. To validate our simulation model and ensure that it provides a realistic benchmark to the baseline (i.e., observed values from the current practice) at the Mayo Clinic,

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

	LOS			Wait		
Group	**Baseline**	**Simulation**	**p-value**	**Baseline**	**Simulation**	**p-value**
All	238	238	0.583	36	36	0.748
ESI=1	192	183	0.139	11	12	0.16
ESI=2	276	274	0.221	24	25	0.155
ESI=3	237	239	0.161	42	41	0.254
ESI=4	153	153	0.748	42	42	0.788
ESI=5	87	85	0.759	37	37	0.888

Table 5 Comparison of average LOS and waiting time (minutes) between the simulated and baseline values.

we perform a series of comparisons by making use of two well-defined metrics of operational performance. Specifically, we focus on average waiting time and LOS and calculate them both for the overall population and for each ESI level. We run the simulation for 10 years and discard the first 3 years as a transient period. We compute the hypothesis test statistic for the two metrics of interest, comparing whether the baseline observed from our data has a different distribution compared to what we obtain from our simulation. As shown in Table 5, all p-values for the differences are large (> 0.5), indicating that the simulated system accurately approximates the current practice of the Mayo Clinic. This can also be seen by noting that the difference between the baseline and simulation in terms of both the overall average LOS and waiting time metrics is less than a minute. Hence, our simulation environment provides a realistic test bed to evaluate the impact of our proposed VPP design compared to both the current VPP design and alternative ED patient flow approaches (e.g., FT and PIT) discussed earlier.

7.2. Mapping the Analytical Model to Real-World Healthcare Systems

To identify the optimal VPP design for our partner hospital using our analytical model (Section 4), we need to specify the values of $\alpha, k_1, \mu_V, \lambda$ (see Table EC.1). To compute α, we use the dependent variable of the proposed ML model (Section 6.2). Specifically, we note that the care of the 10,144 patients out of the 49,350 did require an ED bed, and thus, we set our baseline α to 20.55%. Using this value, we next leverage Equation (6), which indicates that $k_1 = 0.175$. We next determine μ_V, which reflects the relative speed of the VPP unit compared to the main ED. By design, physicians who serve patients in the VPP unit strive to complete the consultation within 20 minutes. Assuming this time constraint on average, we can focus only on the average service rate of the main ED per hour of the day and make use of it to obtain the ratio between μ_V and μ_E. Our analysis shows that, at our partner hospital, the service rate of the VPP unit is six to eight times that of the main

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

ED. Thus, on average, the service duration in the main ED is between 120 minutes and 160 minutes, depending on the hour of the day (see Figure EC.2a). From Proposition 2, we next compute the values of μ_2 and μ_4, verifying that $\mu_2 < \mu_V < \mu_4$. Following Table 1, we observe that our partner hospital falls in the policy regime where $\tau^* = \alpha$ for all $0 < \lambda < \lambda_3$ (Figure EC.2b). Figure EC.2 illustrates the sensitivity analysis on the system parameters to ensure that the proposed policy is robust to data perturbations and hourly changes during the day. We also repeat our analysis for another hospital, Boston Medical Center (BMC), using their data to create yet another benchmark and show how the optimal VPP routing policy depends on the hospital characteristics (see Section EC.2 for analysis related to BMC).

7.3. Combined Routing and Patient Prioritization

Our analyses in the previous section shed light on the best policies that should be followed in practice for routing patients to the VPP unit. However, among patients that are routed to the VPP unit, an ED can follow various prioritization mechanisms. Augmenting patient routing policies with prioritization rules might yield significant benefits in practice. To gain insights into suitable rules that allow for both routing and prioritization, we consider three implementable policies and compare them with the current practice at the Mayo Clinic. Specifically, we consider the following policies:

- **Baseline**: This scenario simulates the current practice at the Mayo Clinic. The implementation of the VPP operation is guided by the empirical data as described in Section 7.1.
- $\boldsymbol{\tau}^*$(**ESI**): Following the design presented in Figure 4, under this policy, all patients with $\hat{\gamma} < \tau^*$ are routed to the VPP, where $\hat{\gamma}$ is obtained from the ML model. Furthermore, among patients with $\hat{\gamma} < \tau^*$, priority is given to patients with a lower ESI level. That is, patients are (a) routed to the VPP unit based on the ML model's score and (b) prioritized there based on their ESI level.
- $\boldsymbol{\tau}^*(\hat{\gamma})$: Similar to the previous policy, patients with $\hat{\gamma} < \tau^*$ are routed to the VPP. However, instead of ESI, prioritization is done based on the predicted score, $\hat{\gamma}$. That is, both routing and priority decisions for utilizing the VPP unit are based on the ML model's output.
- **ESI**: Under this policy, both routing and priority decisions are based on the ESI level. In particular, all patients with ESI>3 (i.e., low acuity patients) are routed to the

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

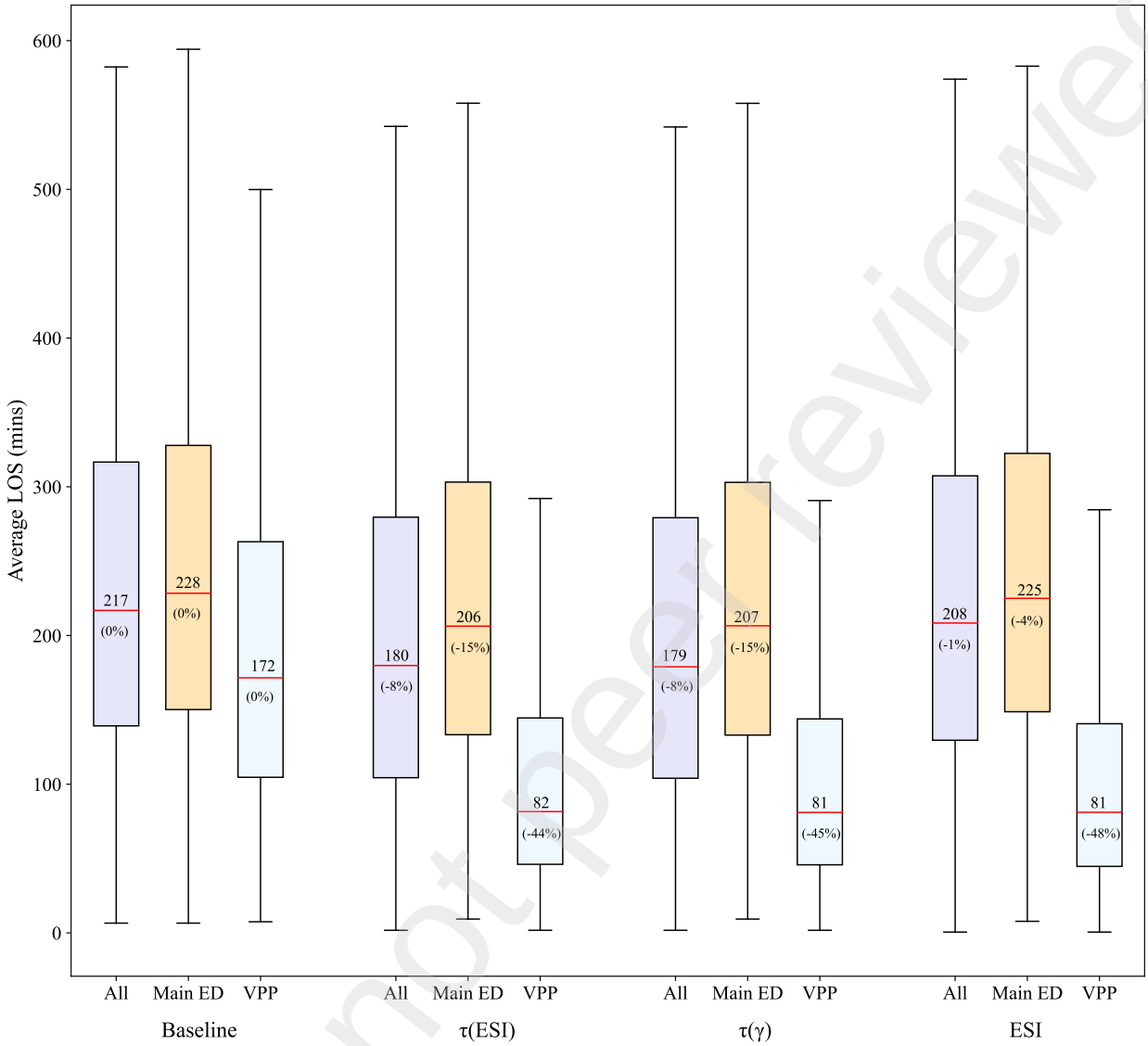


Figure 10 **Average LOS of all patients, patients served in the main ED, and patients served in the VPP unit at the Mayo Clinic ED (% reduction over the baseline is indicated in parentheses).**

VPP unit. Under this policy, we assume that the ML model is not implemented, and instead, a strict rule based on ESI is used (similar to how EDs make use of their FT units).

Figure 10 shows that all of the three policies considered ($\boldsymbol{\tau}^*(\mathbf{ESI})$, $\boldsymbol{\tau}^*(\hat{\boldsymbol{\gamma}})$, **ESI**) lead to substantial improvements in the overall system's performance compared to current practice. This is to some extent expected, given that in the current practice at our partner hospital VPP routing and priority decisions are made in an ad-hoc manner by individual physicians. Furthermore, we observe that the $\boldsymbol{\tau}^*(\hat{\boldsymbol{\gamma}})$ policy results in an average LOS of 202.9 minutes, which corresponds to a 8.0% reduction compared to the current practice.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

We observe small differences between $\boldsymbol{\tau}^*(\hat{\boldsymbol{\gamma}})$ and $\boldsymbol{\tau}^*(\mathbf{ESI})$. This is mainly because ESI is the primary driver of risk for $\hat{\gamma}$ (see the SHAP graph in Figure 9). Hence, there are only minor differences between these two prioritization policies. However, we observe that both of these lead to significant benefits compared to the **ESI** policy, highlighting that using the ML model and following a data-driven VPP design is superior to an ESI-based rule that blindly sends the low acuity (ESI>3) patients to the VPP unit. Similar findings are uncovered when we focus on the average waiting time in the system. The $\boldsymbol{\tau}^*(\hat{\boldsymbol{\gamma}})$ policy, for example, improves the average patient waiting time by 48.0%.

Put together, these results indicate that our partner hospital should change the current practice of routing patients to the VPP unit. In particular, we find that making use of the ML model to obtain predicted risk scores and following the $\boldsymbol{\tau}^*(\hat{\boldsymbol{\gamma}})$ policy can go a long way. Our analysis for the BMC ED also leads to a similar conclusion, suggesting that the $\boldsymbol{\tau}^*(\hat{\boldsymbol{\gamma}})$ policy yields the greatest overall reduction of LOS in the system (see Figure EC.3).

8. VPP or Other Flow Designs: What Hospitals Should Introduce a VPP Unit?

In the previous section, we answered the first two research questions we raised in Section 1. Specifically, we found that an ML model can be developed to reliably identify patients who do not need an ED bed, and that the best way of utilizing the VPP unit is to follow the $\tau^*(\hat{\gamma})$ policy, in which both routing and prioritization decisions are made using the ML model's output as well as the main ED characteristics. In this section, we turn to our third research question: for what hospitals the best VPP design outperforms other ED flow designs such as FT and PIT? To address this question, we simulate counterfactual non-VPP designs, including FT-based and PIT-based streaming approaches introduced in Section 1 (see Section EC.3 for more details about the assumptions we make to simulate performance under these counterfactual designs). Furthermore, since the population of patients served by an ED differs from one hospital to another, we also conduct a sensitivity analysis on the main characteristics of the patient population served by the ED, which in turn enables us to answer our third research question.

We evaluate the impact of FT, PIT, and VPP (under the best policy, $\tau^*(\hat{\gamma})$) on the resulting LOS across all patients served, patients served only in the FT/VPP, and patients served in the main ED. In addition, to generalize our insights beyond the context of our

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

Population	Mean ESI	FT	PIT	VPP
All	2.39	274.7 (273.6, 275.8)	Not Stable	258.8* (257.7, 259.8)
	2.76	233.6 (232.7, 234.4)	1277.0 (1253.5, 1300.5)	203.7* (202.8, 204.6)
	3.03	Not Stable	260.9 (259.8, 261.9)	208.0* (206.6, 209.4)
	3.37	Not Stable	174.0* (173.5, 174.5)	Not Stable
	Mayo ED	232.9 (232.1, 233.8)	785.1 (773.8, 796.4)	202.9* (202.0, 203.7)
Main ED	2.39	288.6 (287.5, 289.7)	Not Stable	272.7* (271.6, 273.7)
	2.76	249.2 (248.2, 250.1)	1419.1 (1392.9, 1445.4)	232.4* (231.4, 233.4)
	3.03	Not Stable	297.2 (296.0, 298.4)	219.9* (218.5, 221.3)
	3.37	Not Stable	198.8* (198.2, 199.3)	Not Stable
	Mayo ED	245.6 (244.7, 246.6)	861.4 (848.8, 873.9)	228.3* (227.3, 229.2)
FT/VPP	2.39	96.6* (95.2, 98.1)	Not Stable	109.9 (107.3, 112.6)
	2.76	164.6 (163.1, 166.2)	114.9 (113.6, 116.3)	103.7* (102.4, 104.9)
	3.03	Not Stable	104.8* (103.9, 105.7)	199.4 (197.3, 201.5)
	3.37	Not Stable	99.1* (98.4, 99.7)	Not Stable
	Mayo ED	173.5 (171.9, 175.2)	124.2 (122.8, 125.7)	108.5* (107.2, 109.9)

Table 6 Average LOS and 95% confidence intervals (indicated in parentheses) per patient subgroup across different ED flow designs. Four synthetic patient populations with varying mean ESI scores are considered in addition to the Mayo Clinic baseline sample. We indicate with an asterisk the best performing system for each population subgroup.

partner institution, we generate synthetic populations of ED patients by altering the distributions of ESI levels and patient's age. We focus on these two factors, mainly because they constitute the two most predictive patient characteristics, as shown in Figure 9, that are associated with the likelihood of requiring an ED bed. Specifically, we split the synthetically generated patient population into distinct groups based on their ESI level (1 to 5) and age ($[0,40),[40,50),[50,60),[60,70),[70,100)$). Subsequently, guided by other ED environments described in the literature, we uniformly sample without replacement from each of the subgroups to generate patient populations of varying severity and care needs that approximate different community profiles that can be served by an ED (Wong et al. 2021, Xu et al. 2009, Araz et al. 2019). In terms of ESI, we let the mean ESI range from 2.39 and 3.37 (see Table 6). In terms of age, we alter the age distribution such that the average patient age lies in the set $\{40,50,60,70\}$ (see Table 7).

Tables 6 and 7 summarize our results. Overall, our analysis shows that given a fixed amount of ED resources, the optimal VPP design outperforms the FT and PIT designs for EDs that serve a patient population with low to medium-high mean ESI scores. However, when the patient body served in the ED has a low prevalence of acute and critical conditions (i.e., involves a low fraction of high ESI level patients), our results suggest that the PIT system is the most suitable design. This is because the flow of patients to the VPP

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

Population	Mean Age	FT	PIT	VPP
All	40	263.3 (262.2, 264.4)	456.7 (453.1, 460.3)	186.9* (186.0, 187.7)
	50	253.1 (252.1, 254.1)	536.9 (532.0, 541.9)	190.3* (189.4, 191.1)
	60	247.4 (246.5, 248.3)	616.8 (610.4, 623.2)	192.9* (192.0, 193.8)
	70	243.1 (242.2, 244.0)	1092.7 (1072.5, 1112.8)	195.6* (194.7, 196.4)
	Mayo ED	232.9 (232.1, 233.8)	785.1 (773.8, 796.4)	202.9* (202.0, 203.7)
Main ED	40	239.4 (238.4, 240.3)	509.2 (505.1, 513.2)	223.6* (222.5, 224.6)
	50	241.1 (240.1, 242.0)	599.4 (593.8, 605.1)	227.1* (226.1, 228.1)
	60	243.1 (242.2, 244.1)	687.2 (680.0, 694.4)	227.7* (226.7, 228.8)
	70	244.2 (243.3, 245.2)	1223.9 (1201.1, 1246.7)	229.3* (228.2, 230.3)
	Mayo ED	245.6 (244.7, 246.6)	861.4 (848.8, 873.9)	228.3* (227.3, 229.2)
FT/VPP	40	332.6 (329.6, 335.7)	115.6 (114.5, 116.8)	106.8* (105.8, 107.9)
	50	290.1 (287.4, 292.8)	115.8 (114.6, 117.0)	105.0* (103.9, 106.0)
	60	261.5 (259.2, 263.9)	115.8 (114.6, 117.1)	105.7* (104.6, 106.8)
	70	239.1 (237.0, 241.2)	117.6 (116.4, 118.9)	106.4* (105.2, 107.5)
	Mayo ED	173.5 (171.9, 175.2)	124.2 (122.8, 125.7)	108.5* (107.2, 109.9)

Table 7 Average LOS and 95% confidence intervals (indicated in parentheses) per patient subgroup across different ED flow designs. Four synthetic patient populations with a varying mean age at admission are considered in addition to the Mayo Clinic baseline sample. We indicate with an asterisk the best performing system for each population subgroup.

unit significantly increases as the patient population shifts toward lower acuity patients, rendering the VPP design unstable. This suggest that the VPP design is more suitable for trauma centers or regular teaching hospital EDs, but the PIT system might be the preferred design in community hospitals where a higher fraction of patients are of low acuity. Finally, as shown in Table 6, we find that the FT approach faces the same problem as the VPP design in EDs with a high fraction of low acuity patients. The simulation outcomes also validate the findings from our analytical model: in cases of very high arrival rates to the ED, all patients should be first routed to the VPP unit, making the VPP and PIT designs similar in their functioning and performance.

Table 7 illustrates that, given a certain set of resources, the age distribution does not impact the ranking of the three design approaches considered. Of note, the average LOS of all patients in the system increases (decreases) when the distribution shifts to older populations in the case of the VPP and PIT (FT). The opposite trend in the FT is driven by the system behavior outside of the main ED. The average LOS of patients served in the FT significantly decreases for older populations, contrary to the case of PIT and VPP where the performance does not significantly change due to age variations. When focusing on the main ED patients, we observe that the LOS measure under PIT significantly increases for older populations. In the case of FT and VPP, we still observe an increase in the average

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

LOS but with a smaller variation. For example, when the mean age is equal to 40, the average LOS in the main ED for the VPP system is 223.6 minutes, for the FT is 239.4 minutes, and for the PIT approach is 509 minutes. When the average age is as high as 70 years, the LOS under PIT increases by 714 minutes while the LOS under VPP and FT increases by only about five minutes. These results highlight that the adaptability of the optimal VPP design leads to lower variations in the system's LOS as the population characteristics change. Specifically, between the 40 and the 70 age groups, the difference in the average performance of the VPP was nine minutes, while in the case of FT and PIT it was 20 minutes and 636 minutes, respectively. Our findings highlight the robustness of the VPP patient flow design (and the potential of a data-driven implementation) that is flexible to changes in various system characteristics, such as population complexity, service rate, and arrival intensity.

9. Discussion and Conclusion

Vertical processing has been popularised as a streaming approach by the Mayo Clinic, even though it has only been implemented in practice in an ad-hoc fashion. To this day, there is a lack of conclusive evidence on the benefit that vertical processing can bring to ED operations. Our investigation proposes, for the first time, a roadmap on how to design, optimize, and implement a VPP-based patient flow in EDs in a data-driven fashion. Our results also shed light on the conditions and hospitals for which implementing a VPP design outperforms other forms of patient streaming.

Since the benefit of using a VPP unit largely depends on accurate up-front predictions of patients that can be served vertically (as opposed to horizontally), we introduced an ML-guided VPP design that makes use of both patient triage information as well as the ED system characteristics to personalize the routing decision for each patient. However, we recognize that the validation of an accurate ML model to determine routing eligibility might not be sufficient for a successful implementation in practice. For this reason, we propose a novel and generalizable analytical queuing model that characterizes the system's optimal policy as a function of ML performance, the patient population characteristics, and the ED's operational load. Our approach considers the patient streaming design in a holistic way, proposing a solution that is not founded on distributional information but rather on individual patient records. Thus, even though the proposed system is fully

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

data-driven, its implementation is guided by a rigorous analytical approach that aims to minimize the operational load of the ED. An important aspect of this analytical approach is that it allows for optimizing ED performance based on three elements: an ED's operational characteristics, the patient population the ED serves, and the predictive power of the ML model the ED wishes to implement.

Our analytical approach shows that there exists a classification threshold based on the patient's predicted complexity below which patients should be routed to the VPP. The value of the threshold determines the volume of cases and, thus, the load in the queue of the main ED vis-a-vis the VPP unit. Our analysis shows that the optimal classification threshold τ^* increases as the overall arrival rate to the ED or the speed of service at the VPP unit increase. Under low to moderate service and arrival rates in the ED, our analysis reveals that the threshold should be set equal to the baseline, namely the overall proportion of the patient population that can be served without an ED bed. This region expands to even busier or faster systems as the discrimination performance of the ML improves, which in turn sheds light on the interplay between operations and the quality of the implemented ML model. Of note, when the ED system lies in scenarios of very high arrival rates and the VPP is significantly faster compared to the ED, our policy suggests a "risk-seeking" behavior in which $\tau^* = 1$. Under this setting, all patients become VPP eligible, and the VPP streaming adapts to a PIT streaming approach. Moreover, our sensitivity analysis across different system parameters reveals that the policy regions under the VPP design are fairly robust to time-dependent variations of the arrivals to the ED. Thus, even though the queueing model is time-invariant the identified optimal policy can be implemented without concerns related to time-dependent spikes in arrivals.

Finally, by developing a data-driven realistic simulation model, we demonstrate that vertical processing can lead to substantial improvements in ED efficiency compared to an FT and PIT approach, reducing the average patient waiting time and LOS in the system. However, our sensitivity analysis suggests that this finding is not universal. Vertical streaming is most beneficial for EDs with a high proportion of patients of high acuity (low ESI scores). In settings where the majority of patients can be treated with a limited amount of tests and resources, the FT has an edge over the VPP approach.

Future research can extend our work by performing a prospective follow-up study of the proposed VPP flow design. In particular, our models need to be externally validated by a

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

different institution prior to implementation using retrospective electronic health records. A potential follow-up study could also aim at prospectively curating a larger and more detailed dataset for establishing VPP eligibility of ED patients.

Despite these limitations, by combining features of the ED, the patient population, and the ML model into one analytical framework, we provide a holistic design for VPP units along with evidence for its superior performance across multiple ED settings. Our results show that the VPP design should be viewed by various EDs as an effective and yet inexpensive solution for enhancing performance. Importantly, VPP implementation only involves the use of a dedicated room with limited physical space and no additional expensive resources (e.g., ED bed or added physician), posing minimal constraints compared to the other forms of patient streaming, As such, we hope to see a broader set of experimentation and potential adoption across various EDs in the near future.

References

Araz OM, Olson D, Ramirez-Nafarrate A (2019) Predictive analytics for hospital admissions from the emergency department using triage information. *International Journal of Production Economics* 208:199–207.

Association AH, et al. (2002) Emergency department overload: A growing crisis. *The results of the American Hospital Association survey of Emergency Department (ED) and hospital capacity. Falls Church, VA: American Hospital Association* 2002.

Benabbas R, Shah R, Zonnoor B, Mehta N, Sinert R (2020) Impact of triage liaison provider on emergency department throughput: A systematic review and meta-analysis. *The American Journal of Emergency Medicine* 38(8):1662–1670.

Bertsimas D, Lukin G, Mingardi L, Nohadani O, Orfanoudaki A, Stellato B, Wiberg H, Gonzalez-Garcia S, Parra-Calderón CL, Robinson K, et al. (2020a) Covid-19 mortality risk assessment: An international multi-center study. *PloS One* 15(12):e0243262.

Bertsimas D, Orfanoudaki A, Weiner RB (2020b) Personalized treatment for coronary artery disease patients: a machine learning approach. *Health Care Management Science* 23:482–506.

Breiman L (2001) Random forests. *Machine Learning* 45(1):5–32.

Breiman L, Friedman JH, Olshen RA, Stone CJ (2017) *Classification and Regression Trees* (Routledge).

Chaou CH, Chen HH, Chang SH, Tang P, Pan SL, Yen AMF, Chiu TF (2017) Predicting length of stay among patients discharged from the emergency department—using an accelerated failure time model. *PloS One* 12(1):e0165756.

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining* 785–794.

Cortes C, Vapnik V (1995) Support-vector networks. *Machine Learning* 20(3):273–297.

Cowan RM, Trzeciak S (2004) Clinical review: emergency department overcrowding and the potential impact on the critically ill. *Critical Care* 9(3):1–5.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

Derlet RW, Richards JR (2002) Emergency department overcrowding in Florida, New York, and Texas. *Southern Medical Journal* 95(8):846–850.

Di Somma S, Paladino L, Vaughan L, Lalle I, Magrini L, Magnanti M (2015) Overcrowding in emergency department: an international issue. *Internal and Emergency Medicine* 10(2):171–175.

Feizi A, Carson A, Jaeker JB, Baker WE (2022) To batch or not to batch? impact of admission batching on emergency department boarding time and physician productivity. *Operations Research* (forthcoming).

Fields WW, Asplin BR, Larkin GL, Marco CA, Johnson LA, Yeh C, Ghezzi KT, Rapp M (2001) The emergency medical treatment and labor act as a federal health care safety net program. *Academic Emergency Medicine* 8(11):1064–1069.

Franklin BJ, Li KY, Somand DM, Kocher KE, Kronick SL, Parekh VI, Goralnick E, Nix AT, Haas NL (2021) Emergency department provider in triage: assessing site-specific rationale, operational feasibility, and financial impact. *Journal of the American College of Emergency Physicians Open* 2(3):e12450.

Gill SD, Lane SE, Sheridan M, Ellis E, Smith D, Stella J (2018) Why do "fast track" patients stay more than four hours in the emergency department? an investigation of factors that predict length of stay. *Emergency Medicine Australasia* 30(5):641–647.

Hastie T, Tibshirani R, Friedman JH, Friedman JH (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer).

Head T, Kumar M, Nahrstaedt H, Louppe G, Shcherbatyi I (2020) Scikit-optimize/scikit-optimize. *(version 0.8. 1)* .

Hodgson NR, Saghafian S, Klanderman MC, Urumov A, Traub SJ (2023) Physician-driven early evaluation: Encounters seen in a vertical model. *JEM Reports* 100028.

Hong WS, Haimovich AD, Taylor RA (2018) Predicting hospital admission at emergency department triage using machine learning. *PloS One* 13(7):e0201016.

Izady N, Mohamed I (2021) A clustered overflow configuration of inpatient beds in hospitals. *Manufacturing & Service Operations Management* 23(1):139–154.

Klug M, Barash Y, Bechler S, Resheff YS, Tron T, Ironi A, Soffer S, Zimlichman E, Klang E (2020) A gradient boosting machine learning model for predicting early mortality in the emergency department triage: devising a nine-point triage score. *Journal of General Internal Medicine* 35(1):220–227.

Lane BH, Mallow PJ, Hooker MB, Hooker E (2020) Trends in united states emergency department visits and associated charges from 2010 to 2016. *The American Journal of Emergency Medicine* 38(8):1576–1581.

Lee SB, Kim DH, Kim T, Kang C, Lee SH, Jeong JH, Kim SC, Park YJ, Lim D (2020) Emergency department triage early warning score (trews) predicts in-hospital mortality in the emergency department. *The American Journal of Emergency Medicine* 38(2):203–210.

Li W, Sun Z, Hong LJ (2021) Who is next: Patient prioritization under emergency department blocking. *Operations Research* (forthcoming).

Lucero A, Sokol K, Hyun J, Pan L, Labha J, Donn E, Kahwaji C, Miller G (2021) Worsening of emergency department length of stay during the covid-19 pandemic. *Journal of the American College of Emergency Physicians Open* 2(3):e12489.

Lundberg S, Lee SI (2017) A Unified Approach to Interpreting Model Predictions. Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds., *Advances in Neural Information Processing Systems 30*, 4765–4774 (Curran Associates, Inc.).

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI (2020) From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2(1):56–67.

Mackenzie Bean (2023) 39 hospitals with the most ed visits. `https://www.beckershospitalreview.com/rankings-and-ratings/hospitals-with-the-most-er-visits.html`.

Moore BJ, Liang L (2020) Costs of emergency department visits in the united states, 2017. *Healthcare Cost and Utilization Project (HCUP) Statistical Briefs [Internet]* (Agency for Healthcare Research and Quality (US)).

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

Partovi SN, Nelson BK, Bryan ED, Walsh MJ (2001) Faculty triage shortens emergency department length of stay. *Academic Emergency Medicine* 8(10):990–995.

Patki N, Wedge R, Veeramachaneni K (2016) The synthetic data vault. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 399–410 (IEEE).

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Raita Y, Goto T, Faridi MK, Brown DF, Camargo CA, Hasegawa K (2019) Emergency department triage prediction of clinical outcomes using machine learning models. *Critical Care* 23(1):1–13.

Rondeau KV, Francescutti LH, Zanardelli JJ (2005) Emergency department overcrowding: the impact of resource scarcity on physician job satisfaction/practitioner application. *Journal of Healthcare Management* 50(5):327.

Rosenblatt F (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* 65(6):386.

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

Saghafian S, Hopp WJ, Iravani SM, Cheng Y, Diermeier D (2018) Workload management in telemedical physician triage and other knowledge-based service systems. *Management Science* 64(11):5180–5197.

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2012) Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research* 60(5):1080–1097.

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2014) Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing & Service Operations Management* 16(3):329–345.

Saghafian S, Kilinc D, Traub S, et al. (2023) *Dynamic Assignment of Patients to Primary and Secondary Inpatient Units: Is Patience a Virtue?* (Cambridge Handbook on Productivity, Efficiency and Effectiveness in Healthcare (forthcoming)).

Schafermeyer RW, Asplin BR (2003) Hospital and emergency department crowding in the United States. *Emergency Medicine* 15(1):22–27.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

Servi LD, Finn SG (2002) $m/m/1$ queues with working vacations ($m/m/1/wv$). *Performance Evaluation* 50(1):41–52.

Sun Y, Heng BH, Tay SY, Seow E (2011) Predicting hospital admissions at emergency department triage using routine administrative data. *Academic Emergency Medicine* 18(8):844–850.

Tang Y (1994) The departure process of the M/G/1 queueing model with server vacation and exhaustive service discipline. *Journal of Applied Probability* 31(4):1070–1082.

Traub SJ, Bartley AC, Smith VD, Didehban R, Lipinski CA, Saghafian S (2016) Physician in triage versus rotational patient assignment. *The Journal of Emergency Medicine* 50(5):784–790.

Traub SJ, Saghafian S, Judson K, Russi C, Madsen B, Cha S, Tolson HC, Sanchez LD, Pines JM (2018) Interphysician differences in emergency department length of stay. *The journal of emergency medicine* 54(5):702–710.

Traub SJ, Wood JP, Kelley J, Nestler DM, Chang YH, Saghafian S, Lipinski CA (2015) Emergency department rapid medical assessment: overall effect and mechanistic considerations. *The Journal of Emergency Medicine* 48(5):620–627.

US Department of Health, Human Services, et al. (2014) The health information technology for economic and clinical health (hitech) act.

Venkatesh AK, Janke A, Rothenberg C, Chan E, Becher RD (2021) National trends in emergency department closures, mergers, and utilization, 2005-2015. *PloS One* 16(5):e0251729.

Wong AH, Whitfill T, Ohuabunwa EC, Ray JM, Dziura JD, Bernstein SL, Taylor RA (2021) Association of race/ethnicity and other demographic characteristics with use of physical restraints in the emergency department. *JAMA Network Open* 4(1):e2035241–e2035241.

Wong HJ, Morra D, Caesar M, Carter MW, Abrams H (2010) Understanding hospital and emergency department congestion: an examination of inpatient admission trends and bed resources. *Canadian Journal of Emergency Medicine* 12(1):18–26.

Xu KT, Nelson BK, Berk S (2009) The changing profile of patients who used emergency department services in the united states: 1996 to 2005. *Annals of Emergency Medicine* 54(6):805–810.

Yoon P, Steiner I, Reinhardt G (2003) Analysis of factors influencing length of stay in the emergency department. *Canadian Journal of Emergency Medicine* 5(3):155–161.

Zhang X, Kim J, Patzer RE, Pitts SR, Patzer A, Schrager JD (2017) Prediction of emergency department hospital admission based on natural language processing and neural networks. *Methods of Information in Medicine* 56(05):377–389.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

Electronic Companion

EC.1. Proofs

All proofs for equations, theorems and lemmas, where applicable, are given below.

Proof of Theorem 1

If $u = 0$, the LOS of the main ED, LOS_E, then becomes the sojourn time of an $M/M/1$ queue since its arrival follows an exponential distribution with rate λ_E, and can be calculated as follows:

$$LOS_E = \frac{1}{1-\lambda_E}.$$

In addition, note that in reality, $\mu_V >> 1$ because the average time spent in the VPP is in the order of ~ 10 minutes, while that of the ED is ~ 200 minutes. Therefore, although the VPP is an $M/M/1$ queue (with no vacation), we can further simplify the VPP by assuming that it is an $M/M/\infty$ queue with a service rate of μ_V. This is essentially the equivalency of an $M/M/1$ queue with an $M/M/\infty$ queue when the rate at which customers arrive is much less than the service rate, and therefore, practically, a queue rarely forms. In this case, the LOS of the VPP, LOS_V, is:

$$LOS_V = \frac{1}{\mu_V}.$$

Hence, the overuse and underuse costs defined in Equations 9 and 10, can be reduced to Equations EC.1 and EC.2, respectively.

$$C_O(\tau \mid k_1, \lambda, \mu) = \frac{1}{\mu_V}. \tag{EC.1}$$

$$C_U(\tau \mid k_1, \lambda, \mu) = \frac{1}{1-(1-\tau+p(\tau)\tau)\lambda} - \frac{1}{\mu_V} \tag{EC.2}$$

Substituting Equations EC.1, EC.2, 7 and 8 in Equation 11 yields the total cost function to be minimized, C_T:

$$C_T(\tau \mid \alpha, k_1, \lambda, \mu_V) =$$

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

$$\begin{cases} -\dfrac{\alpha\lambda+\mu_V}{\lambda\mu_V}+\dfrac{\tau}{\mu_V}+\dfrac{1-(1-\alpha)\lambda}{\lambda\Big(1-\lambda+(1-k_1)\lambda\tau\Big)}, & \text{if } \tau\le\alpha \\ \dfrac{A+B\tau+C\tau^2}{\mu_V\Big(-1+\alpha+\lambda+\alpha\lambda(-2+\alpha+k_1-k_1\tau)\Big)}, & \text{if } \tau>\alpha \end{cases} \tag{EC.3}$$

where:

$$\begin{aligned} A &= -\alpha\Big(-1+\alpha+\lambda+\alpha(-2+\alpha+k_1)\lambda+k_1\mu_V\Big) \\ B &= \Big(-1+\alpha+\lambda+\alpha(-2+\alpha+k_1+\alpha k_1)\lambda+\alpha k_1\mu_V\Big) \\ C &= -\alpha k_1\lambda \end{aligned}$$

The τ that minimizes Equation EC.3 is the optimal fraction of patients that must be routed to the VPP.

We begin by ensuring that C_T is convex for all combinations of $(\alpha, k_1, \lambda, \mu_V)$ in their allowable range. We verify that:

- C_T is continuous at $\tau=\alpha$;
- $\partial^2 C_T/\partial\tau^2 \ge 0$ in both $\tau\le\alpha$ and $\tau>\alpha$ (i.e., second-order condition).

Thus, the function is continuous and convex for all parameters. Next, we find the first-order condition (FOC) that minimizes C_T. We do this separately for $\tau\le\alpha$ and $\tau>\alpha$. For notational convenience, we define:

$$C_1(\tau\mid\alpha,k_1,\lambda,\mu_V)=-\frac{\alpha\lambda+\mu_V}{\lambda\mu_V}+\frac{\tau}{\mu_V}+\frac{1-(1-\alpha)\lambda}{\lambda\Big(1-\lambda+(1-k_1)\lambda\tau\Big)}, \quad \tau\le\alpha$$

$$C_2(\tau\mid\alpha,k_1,\lambda,\mu_V)=\frac{A+B\tau+C\tau^2}{\mu_V\Big(-1+\alpha+\lambda+\alpha\lambda(-2+\alpha+k_1-k_1\tau)\Big)}, \quad \tau>\alpha$$

Case 1: $\boldsymbol{\tau\le\alpha}$

Setting $\partial C_1/\partial\tau=0$, we obtain a unique solution for that minimizes C_1:

$$\tau_1^*=\frac{1-\lambda}{(-1+k_1)\lambda}+\sqrt{\frac{(-1+\lambda-\alpha\lambda)\mu_V}{(-1+k_1)\lambda^2}}, \tag{EC.4}$$

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

where τ_1^* exists when the following conditions hold:

$$0 < \alpha < \frac{1}{2} \quad \text{(for all } \alpha\text{'s)}$$
$$\frac{1}{2} < k_1 < 1 - \alpha$$
$$2 < \mu_V < \mu_1$$
$$\lambda_1 < \lambda < \lambda_2,$$

where μ_1 and λ_1 are defined in Proposition 3 and λ_2 is defined as:

$$\lambda_2(\alpha, k_1, \mu_V) = \frac{2 - \alpha(-1+k_1)(-2+\mu_V) + (-1+k_1)\mu_V}{2\Big(1 - \alpha(1-k_1)\Big)^2} - \frac{\sqrt{(-1+k_1)\mu_V\Big(4\alpha k_1(-1+\alpha-\alpha k_1) + (1-\alpha)^2(-1+k_1)\mu_V\Big)}}{2\Big(1-\alpha(1-k_1)\Big)^2}$$

Outside of the range where the FOC has a unique solution, τ_1^* is either 0 or α, which we determine based on whether $\partial C_1/\partial\tau$ is positive or negative. This algebra yields the following boundary solutions when $0 \le \tau \le \alpha$:

$$\tau_1^* = \begin{cases} \alpha, & \text{if} \quad 0 < k_1 < \frac{1}{2}, \quad \text{since} \quad \partial C_1/\partial\tau \le 0 \\ \alpha, & \text{if} \quad \mu_V > \mu_1, \quad \text{since} \quad \partial C_1/\partial\tau \le 0 \\ 0, & \text{if} \quad 0 < \lambda < \lambda_1, \quad \text{since} \quad \partial C_1/\partial\tau \ge 0 \\ \alpha, & \text{if} \quad \lambda_2 < \lambda < 1, \quad \text{since} \quad \partial C_1/\partial\tau \le 0 \end{cases} \tag{EC.5}$$

Case 2: $\boldsymbol{\tau \ge \alpha}$

Setting $\partial C_2/\partial\tau = 0$, we obtain

$$\tau_2^* = \frac{-1+\alpha+\lambda+\alpha\lambda(-2+\alpha+k_1+k_1\sqrt{-\frac{(-1+\alpha)(1+(-1+\alpha)\lambda)\mu_V}{\alpha k_1\lambda^2}})}{\alpha k_1 \lambda}, \tag{EC.6}$$

where τ_2^* exists when the following conditions hold:

$$0 < \alpha < \frac{1}{2},$$

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

$$0 < k_1 < 1 - \alpha,$$
$$\max\{2, \mu_2\} < \mu_V < \mu_4,$$
$$\lambda_3 < \lambda < \lambda_4;$$

where, $\lambda_3, \lambda_4, \mu_2, \mu_4, k_A$ are defined as:

$$\lambda_3(k_1, \alpha, \mu_V) = \frac{1}{2}\left(\frac{2 - \alpha(2 + k_1(-2 + \mu_V))}{(1 + \alpha(-1 + k_1))^2} - \sqrt{-\frac{\alpha^2 k_1^2 \mu_V (4 + \alpha(-4 + 4k_1 - \mu_V) + \mu_V)}{(-1 + \alpha)(1 + \alpha(-1 + k_1))^4}}\right),$$
$$\lambda_4(\alpha, k_1, \mu_V) = \frac{1 - \alpha - \alpha k_1 \mu_V}{(-1 + \alpha)^2},$$
$$\mu_2(\alpha, k_1) = \frac{1 - \alpha}{k_1},$$
$$\mu_4(\alpha, k_1) = \frac{1 - \alpha}{\alpha k_1}$$
$$k_A(\alpha) = \frac{(2 - \alpha) - \sqrt{3 - 2\alpha}}{1 - \alpha}$$

Outside of the range where the FOC has a unique solution, τ_2^* is either 1 or α, which we determine based on whether $\partial C_2/\partial\tau$ is positive or negative. This algebra yields the following boundary solutions when $\alpha \le \tau \le 1$:

$$\tau_2^* = \begin{cases} \alpha, & \text{if} \quad \lambda < \lambda_3 \quad \text{since} \quad \partial C_2/\partial\tau \ge 0 \\ \alpha, & \text{if} \quad k_1 < k_A \text{ and } \mu_V < \mu_2, \quad \text{since} \quad \partial C_2/\partial\tau \ge 0 \\ 1, & \text{if} \quad \max\{2, \mu_3\} < \mu_V \text{ and } \max\{0, \lambda_4\} < \lambda < 1 \quad \text{since} \quad \partial C_2/\partial\tau \le 0 \end{cases} \tag{EC.7}$$

where, μ_3 and k_A are defined as:

$$\mu_3(\alpha, k_1) = \frac{1 - \alpha}{k_1}$$
$$k_A(\alpha) = -\sqrt{\frac{3 - 2\alpha}{(-1 + \alpha)^2}} + \frac{-2 + \alpha}{-1 + \alpha}.$$

Note that the region $\mu_2 < \mu_V < \mu_3$ and $\lambda > \lambda_4$ does not exist. Therefore, for the sake of simplicity we do not further break down the state space in the remainder of the proof.

With the optimal τ obtained when $0 < \tau < \alpha$ or $\alpha < \tau < 1$, we finally merge the regions to find τ^* for each combination of parameters.

We realize the following relationship:

$$0 < k_A < \frac{1}{2} < 1 - \alpha.$$

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

Note that when $k_1 < \frac{1}{2}$, $\tau_1^* = \alpha$ and C_1 is decreasing. Therefore, $\tau^* = \tau_2^*$ from Case 2.

When $k_1 \geq \frac{1}{2}$, we observe:

$$2 < \mu_2 < \mu_1 < \mu_4.$$

Note from Equation EC.5 that when $k_1 \geq \frac{1}{2}$ and $\mu_V > \mu_1$, $\tau_1^* = \alpha$ and C_1 is decreasing. Therefore, again, $\tau^* = \tau_2^*$ from Case 2.

However, when $k_1 \geq \frac{1}{2}$ and $\mu_V < \mu_1$, the solutions from Case 1 can be the overall solution to τ^*. Observe that the following relationship holds when $\mu_V < \mu_1$:

$$0 < \lambda_1 < \lambda_2 < \lambda_3 < \lambda_4$$

Further, note that when $\lambda_2 < \lambda$, $\tau_1^* = \alpha$ and C_1 is decreasing. Therefore, again, $\tau^* = \tau_2^*$ from Case 2. Also, note that when $\lambda < \lambda_3$, $\tau_2^* = \alpha$ and C_2 is increasing; therefore, $\tau^* = \tau_1^*$ in this case. Hence, overall, when $\lambda < \lambda_2$, $\tau^* = \tau_1^*$.

Table **??** summarizes the optimal threshold τ^* for all parameter combinations.

□

EC.1.1. *Proof of Proposition 1*

To prove Proposition 1 we must show that the area for which $\tau^* = \alpha$ is decreasing with k_1.

For this, it suffices to show that the μ_V and λ ranges in which $\tau^* = \alpha$ are both decreasing with k_1. Referring to Table **??**, these ranges can be readily found. The following statement is true, and thus proves Proposition 1.

For all $0 < \alpha < \frac{1}{2}$, $0 < \lambda < 1$, $0 < k_1 < 1 - \alpha$, $\mu_V > 2$:

$$\begin{cases} \frac{\partial\left(\mu_4(k_1,\alpha) - \mu_2(k_1,\alpha)\right)}{\partial k_1} < 0 \\ \frac{\partial\left(\mu_4(k_1,\alpha) - \mu_1(k_1,\alpha)\right)}{\partial k_1} < 0 \\ \frac{\partial \lambda_3(k_1,\alpha,\mu_V)}{\partial k_1} < 0 \\ \frac{\partial\left(\lambda_3(k_1,\alpha,\mu_V) - \lambda_2(k_1,\alpha,\mu_V)\right)}{\partial k_1} < 0 \end{cases} \tag{EC.8}$$

□

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

k_1	μ_V	λ	τ^*
$k_1 < k_A$	$2 < \mu_V < \mu_2$	$0 < \lambda < 1$	α
	$\mu_2 < \mu_V < \mu_4$	$0 < \lambda < \lambda_3$	α
		$\lambda_3 < \lambda < \lambda_4$	τ_2
		$\lambda_4 < \lambda < 1$	1
	$\mu_4 < \mu_V$	$0 < \lambda < 1$	1
$k_A < k_1 < 1/2$	$2 < \mu_V < \mu_4$	$0 < \lambda < \lambda_3$	α
		$\lambda_3 < \lambda < \lambda_4$	τ_2
		$\lambda_4 < \lambda < 1$	1
	$\mu_4 < \mu_V$	$0 < \lambda < 1$	1
$1/2 < k_1 < 1 - \alpha$	$2 < \mu_V < \mu_1$	$0 < \lambda < \lambda_1$	0
		$\lambda_1 < \lambda < \lambda_2$	τ_1
		$\lambda_2 < \lambda < \lambda_3$	α
		$\lambda_3 < \lambda < \lambda_4$	τ_2
		$\lambda_4 < \lambda < 1$	1
	$\mu_1 < \mu_V < \mu_4$	$0 < \lambda < \lambda_3$	α
		$\lambda_3 < \lambda < \lambda_4$	τ_2
		$\lambda_4 < \lambda < 1$	1
	$\mu_4 < \mu_V$	$0 < \lambda < 1$	1

Table EC.1 Optimal threshold τ^* for all parameter combinations.

Proof of Lemma 3

For each region in Table **??**, note that when $\tau* = 0$ or $\tau* = \alpha$ or $\tau* = 1$, $\dfrac{\partial \tau^*}{\partial \mu_V} = 0$ and $\dfrac{\partial \tau^*}{\partial \lambda} = 0$. For regions where $\tau* = \tau_1$ or $\tau* = \tau_2$, it can also easily be shown that:

$$\begin{cases} \dfrac{\partial \tau_2}{\partial \mu_V} > 0 \\ \dfrac{\partial \tau_2}{\partial \lambda} > 0 \\ \dfrac{\partial \tau_1}{\partial \mu_V} > 0 \\ \dfrac{\partial \tau_1}{\partial \lambda} > 0 \end{cases}$$

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

Also recall that when $k_1 < 1/2$: $0 < \lambda_3 < lambda_4$ and $2 \leq \mu_2 < \mu_4$; and when $1/2 < k_1 < 1-\alpha$: $0 < \lambda_1 < \lambda_2 < \lambda_3 < lambda_4$ and $2 \leq \mu_1 < \mu_4$ so the regions where $\tau^* = 0 < \tau_1 < \alpha < \tau_2 < 1$ are also increasing in μ_V and λ.

□

Proof of Proposition 2

Proposition 2 can be readily inferred from Table **??**. □

Proof of Proposition 3

Proposition 3 can be readily inferred from Table **??**. □

Proof of Lemma 1

Denote the cumulative distribution function (CDF) of departures from the VPP and direct arrivals to the ED (i.e., after triage) by $F_a(t)$ by $F_b(t)$, respectively. The probability of an arrival at time $T \leq t$ can be written as:

$$F_{a\cup b}(t) = \Pr\{T \leq t\} = \Pr\{\text{departure from VPP} < t\} \cup \Pr\{\text{direct arrival} \leq t\} = F_a(t) + F_b(t) - F_a(t)F_b(t)$$

The direct arrival interarrival time distribution is an exponential distribution with rate $(1-\tau)\lambda$. Hence,

$$F_b(t) = 1 - e^{-(1-\tau)\lambda t}$$

To find $F_a(t)$, we leverage the results from Tang (1994). Denote $\lambda_v = p(\tau)\tau\lambda$. We have:

$$\text{Vacation length CDF} = V(t) = 1 - e^{-t/u}$$

$$v(\lambda_v) = \int_0^\infty e^{-\lambda_v x} dV(x) = \frac{1}{1+\lambda_v u}$$

$$\widetilde{V}(t) = \frac{\int_0^\infty V(t+x)\lambda_v e^{-\lambda_v x} dx - v(\lambda_v)}{1 - v(\lambda_v)} = 1 - e^{-t/u}$$

$$p_0 = \frac{(1-\lambda/\mu_V)\Big(1 - v(\lambda_v)\Big)}{\lambda_v u}$$

$$F(t) = 1 - e^{-\lambda_v t}$$

$$G(t) = 1 - e^{-\mu_V t}$$

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

Finally, Equation 21 from Tang (1994) shows the interdeparture time CDF of the VPP in steady state:

$$F_a(t) = (1-p_0)G(t) + p_0 \int_0^t dF(x) * dG(x) * d\widetilde{V}(t)$$

Finally, $f_a(t) = dF_{a\cup b}(t)/dt$

□

Proof of Lemma 2

It can be shown that the functional form of the receiver operating characteristic (ROC) curve of Equation 4 is as follows:

$$TPR(FPR \mid k_1, \alpha) = \begin{cases} \dfrac{(1-\alpha-\alpha k_1)}{(1-\alpha)k_1} FPR, & \text{if } 0 \le FPR < k_1 \\ \\ \dfrac{(1-\alpha-k_1)+\alpha k_1 FPR}{(1-\alpha)(1-k_1)}, & \text{Otherwise.} \end{cases} \tag{EC.9}$$

where, TPR and FPR are the true positive rate and false positive rate, respectively.

The AUC is calculated by integrating the ROC curve, which results in the following:

$$AUC = \int_0^1 TPR(FPR \mid k_1, \alpha) dFPR = 1 - \frac{k_1}{2(1-\alpha)} \tag{EC.10}$$

□

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

EC.2. Mapping the Analytical Model to the Boston Medical Center

In this section, we illustrate how to apply the analytical model to the BMC ED, showcasing the policy variations in a system that serves a higher volume of patients on a daily basis. BMC primarily serves a higher portion of underprivileged population compared to the Mayo ED with greater racial diversity (Bertsimas et al. 2020b). In addition, the proportion of patients that require an ED bed is low due to the high prevalence of low acuity cases. Leveraging the findings of Feizi et al. (2022), we approximate BMC's α by scaling Mayo's α by the ratio of ESI-4 and ESI-5 patients served in BMC to that of Mayo clinic. We also approximate the hourly arrival rates by scaling up Mayo's hourly arrival rate by the ratio of annual patient volume at BMC to that of Mayo (Mackenzie Bean 2023). We further assume that the trained ML model achieves the same performance as the one presented in Section 6.3. As shown in Figure EC.2c, we are in the regime where $k_A > k_1$ (similar to the Mayo ED). Moreover, if we hypothesize that BMC was able to provide an equivalent amount of resources (rooms and physicians) to achieve the same μ_V as the Mayo ED, then $\mu_2 < \mu_V < \mu_4$ would still apply. However, the average arrival rate λ_{BMC}, changes the optimal regime throughout the day. Specifically, as illustrated in Figure EC.2b, $0 < \lambda < \lambda_3$ between 4.00 pm and 8.00 am ($\tau^* = \alpha$); $\lambda_3 < \lambda < \lambda_4$ between 8.00 am and 9.00 am as well as between 3.00 pm and 4.00 pm. ($\tau^* = \tau_2$); $\lambda_4 < \lambda$ during the hours of 9.00am and 4.00pm ($\tau^* = 1$). This setting highlights the potential variability of the optimal policy throughout the day. In practice, the ED administrators of the BMC, could approximate the optimal design by implementing $\tau^* = \alpha$ during the hours of low demand and increasing it to $\tau^* = 1$ throughout the morning and afternoon hours, leveraging the VPP as a screening tool for any patient in the ED.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

EC.3. Simulated Design for the FT and PIT Systems

We design the FT and PIT based on Feizi et al. (2022) and Franklin et al. (2021), respectively, which employ these policies in their study settings. Below we provide details on the implementation of the FT and PIT policies:

- **Fast-Track (FT)**: All patients with ESI>3 are routed to dedicated beds in the FT section of the department while only patients with ESI≤ 3 use the resources available in the main ED. We assume that at any hour, the FT is staffed with half as many physicians in the main ED and that one patient at a time can be served by an FT worker.
- **Physician-In-Triage (PIT)**: All patients are first seen by a physician during the triage stage, and only patients who require ED care will be sent to the queue. Triage physicians may also initiate the tests. In implementing the PIT policy, we assume that there are always two physicians at the triage stage and that the examination time of a physician is similar to that of an VPP. However, the main ED will operate with two fewer physicians during the hours in which it was originally staffed with over two physicians.

Our analysis attempts to match all three patient streaming systems in terms of the number of resources (i.e., the total number of beds and physicians) throughout a simulated day. Thus, it is possible, under specific conditions, to study scenarios under which at least one of the approaches leads to an unstable system (see Table 6). Note that inevitably we must have two beds working simultaneously and additional physicians in simulating the FT since, by design, it must operate with two separate sections (FT and main ED).

To perform a realistic comparison across the VPP, FT, and PIT approaches, we leverage the synthetically generated data from the Mayo Clinic.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

EC.4. Additional Figures and Tables

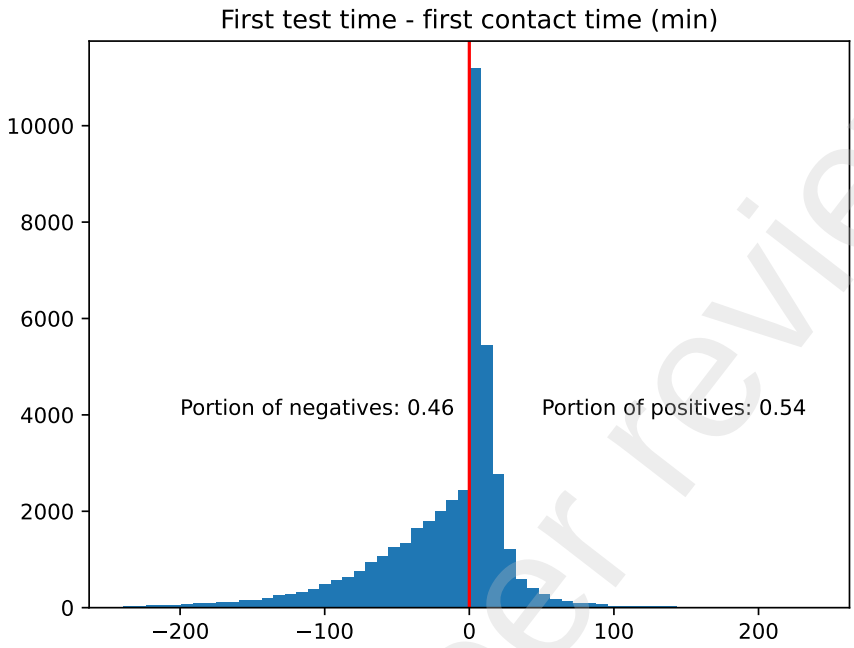


Figure EC.1 **Time from first contact to first test. Negative values indicate that test was ordered prior to contact with physician.**

Preprint not peer reviewed

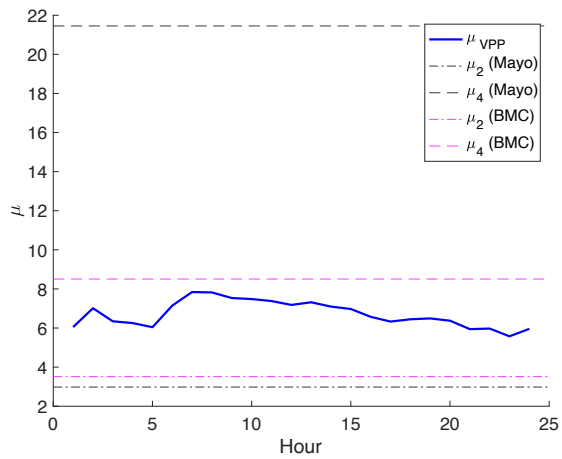
This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

Hour of the Day	ESI=1	ESI=2	ESI=3	ESI=4	ESI=5
0	0.00%	0.00%	0.00%	0.78%	0.00%
1	0.00%	0.50%	0.00%	0.00%	0.00%
2	0.00%	0.00%	0.00%	0.00%	0.00%
3	0.00%	0.00%	0.00%	0.00%	0.00%
4	0.00%	0.00%	0.00%	0.00%	
5	0.00%	0.00%	0.00%	0.00%	0.00%
6	0.00%	0.00%	0.00%	0.00%	0.00%
7	0.00%	0.00%	0.00%	0.00%	0.00%
8	0.00%	0.00%	0.31%	0.97%	0.00%
9	0.00%	0.11%	1.01%	2.40%	13.33%
10	0.00%	2.09%	4.95%	10.41%	15.79%
11	0.00%	4.46%	12.14%	21.05%	26.09%
12	0.00%	5.16%	17.20%	24.59%	68.75%
13	0.00%	5.00%	16.17%	28.57%	33.33%
14	0.00%	5.83%	15.05%	24.07%	31.25%
15	0.00%	3.86%	13.82%	20.06%	30.00%
16	0.00%	5.70%	11.92%	17.88%	40.00%
17	0.00%	3.98%	11.00%	13.83%	10.53%
18	0.00%	1.68%	6.64%	9.60%	18.18%
19	0.00%	1.54%	2.07%	6.19%	16.67%
20	0.00%	0.94%	1.05%	1.48%	4.55%
21	0.00%	0.20%	0.69%	1.15%	6.25%
22	0.00%	0.00%	0.39%	0.51%	0.00%
23	0.00%	0.33%	0.00%	0.00%	0.00%

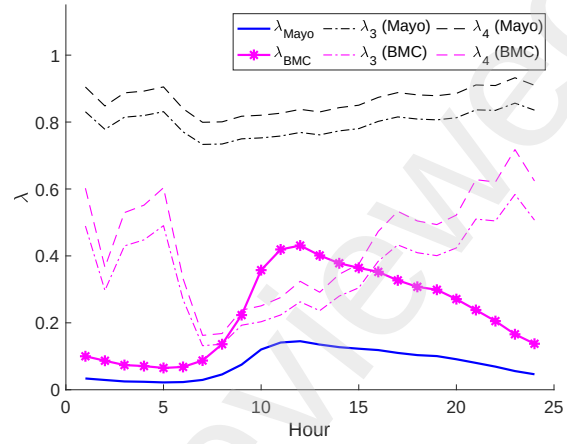
Table EC.2 Proportion of patients served in the VPP of the Mayo ED during the study period for each ESI level.

Preprint not peer reviewed

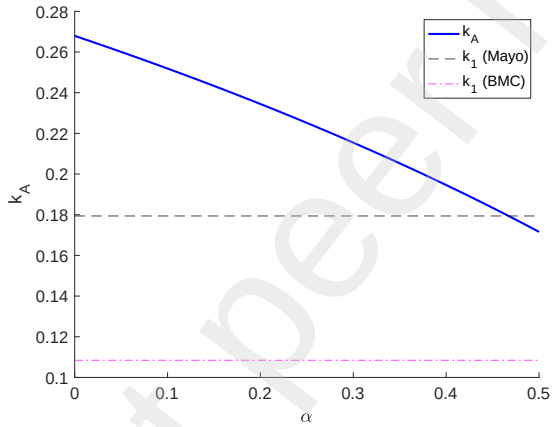
This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161



(a) μ



(b) λ



(c) k_1

Figure EC.2 Sensitivity analysis of the Mayo Clinic ED system parameters.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161

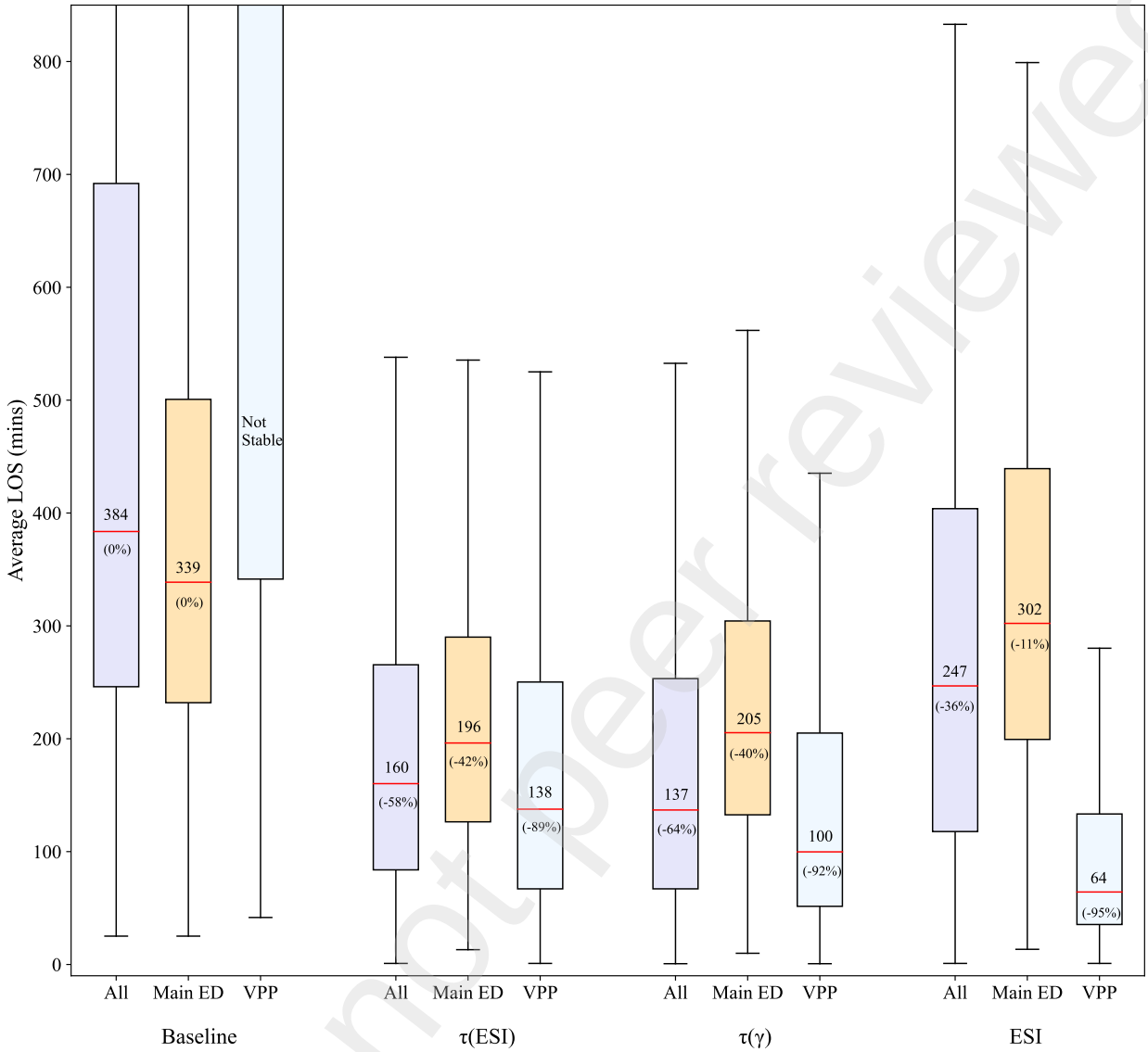


Figure EC.3 **Average LOS of all patients, patients served in the main ED and patients served in the VPP at the BMC ED (the % reduction over the baseline is indicated in parentheses).**

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=4465161