Proactive Transfer Admission Control for Emergency Departments

Ruicheng Ao[1] Jing Dong[2] Xiaole (Alyssa) Liu[3] Martin S. Copenhaver[4]
[1]Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA 02139, aorc@mit.edu
[2]Columbia University, New York, NY 10027,jing.dong@gsb.columbia.edu
[3]Stern School of Business, New York University, New York, NY 10012, xl2500@stern.nyu.edu
[4] School of Medicine, Johns Hopkins University, Baltimore, MD 21287, copenhaver@jhmi.edu

Emergency Department (ED) crowding remains a persistent challenge, undermining timely care, elevating clinical risk, and placing substantial strain on hospital resources. While hospitals cannot easily control the arrival of walk-in emergency patients, they often have some flexibility in managing transfer admissions from other hospitals. Temporarily restricting such transfers during periods of high congestion can provide an effective operational lever to alleviate crowding. In this work, we investigate how predictive information, namely forecasts of future demand and inpatient admission delays, can be systematically incorporated into ED admission control decisions. We develop a queueing model that captures key patient flow dynamics, accounting for a time-varying operating environment, two distinct sources of ED arrivals, and delayed inpatient admissions (i.e., ED boarding). Building on this model, we derive the optimal admission control policy under a fluid approximation. The fluid-based policy can be seamlessly adapted to the stochastic system. It explicitly leverages predictive information, offers clear interpretability and practical implementability, and achieves performance close to the stochastic optimum. Through extensive numerical experiments and a case study based on real hospital data, we show that by proactively regulating the acceptance of transfer patients based on real-time congestion levels and anticipated future conditions, the proposed approach achieves a superior efficiency frontier balancing ED congestion and transfer acceptance than reactive admission policies.

Key words: emergency department, admission control, demand surge, queueing theory, asymptotic analysis

1. Introduction

Emergency department (ED) crowding has emerged as one of the most pervasive and persistent challenges in hospital care delivery. It reflects not only localized operational bottlenecks but also broader systemic strain, and has been shown to adversely affect patient outcomes and contribute to provider burnout (Hoot and Aronsky 2008, Sun et al. 2013). Despite widespread recognition of the problem, hospitals face substantial constraints in expanding capacity. Staffing shortages, rising costs, and the limited availability of physical space make it increasingly difficult for hospitals to simply "scale up" their way out of crowding (Janke and Venkatesh 2025).

Given these barriers, there is growing interest in operational strategies that manage, not just accommodate, demand. While hospitals have little control over unscheduled walk-in emergency

arrivals, many have at least some flexibility in how they manage inter-facility transfer requests which are subject to certain regulations such as the Emergency Medical Treatment And Labor Act (EMTALA) in the United States (Centers for Medicare and Medicaid Services 2026). These transfers play an important role in regionalized care systems, ensuring that patients receive the appropriate level of care and clinical services (Westfall 2024), but they also constitute a modifiable stream of incoming patients whose timing and volume can meaningfully influence ED congestion (Greenwood-Ericksen et al. 2025). Temporarily regulating such admissions during periods of high crowding offers a promising lever for improving patient flow. In this work, we examine how to design effective admission control policies for inter-facility transfer patients (hereafter, transfer patients) to help alleviate ED crowding.

Importantly, ED crowding is not merely an ED-level operational issue; it reflects a system-wide patient flow problem. A substantial portion of ED congestion arises from downstream bottlenecks, particularly delays in admitting patients to inpatient units (McKenna et al. 2019). When inpatient beds are unavailable or admission processing is slow, admitted patients remain boarded in the ED, consuming space and staff resources and exacerbating crowding. Therefore, any meaningful approach to ED admission control must account not only for current ED conditions but also for inpatient-side congestion and the anticipated delays in admitting patients.

At the same time, recent advances in predictive analytics, combined with the growing availability of granular operational data, have created new opportunities for hospitals to anticipate rather than react to congestion. Forecasting models for ED arrivals and inpatient admission delays have become increasingly accurate and operationally feasible for real-time deployment. These developments raise a natural question: How can predictive information be systematically incorporated into admission control policies for transfer patients? In this work, we address this question by developing a modeling framework that integrates forecasts of future ED demands and inpatient admission delays into the design of admission control policies.

To study this problem, we develop a queueing model that captures the essential dynamics of ED patient flow to support the design of effective admission control policies. Our model distinguishes between two streams of arrivals: (1) *Regular Arrivals*, including walk-ins, ambulance arrivals, and referrals from external providers; and (2) *Transfer Requests*, representing patients seeking transfer from other EDs. Admission control is applied exclusively to transfer requests. The model also explicitly incorporates stochastic inpatient admission delays, represented through boarding times that arise from limited inpatient capacity and other operational constraints. By accounting for the time-varying nature of ED operations and these key system features, the model provides a

flexible framework for analyzing policies that balance patient waiting costs with the blocking costs associated with rejecting transfer requests.

We then analyze a deterministic, continuous-time fluid approximation of the system and derive the corresponding optimal control policy, which explicitly incorporates forecasts of future arrival and transfer processing rates. The fluid-based policy can be seamlessly adapted to the stochastic system, maintaining an intuitive structure and ease of implementation while achieving near-optimal performance. Notably, under certain regularity conditions, we rigorously establish the asymptotic optimality of the proposed policy in the many-server limit. The policy operates by dynamically adjusting admission decisions for transfer requests based on real-time system congestion, predicted future demand, and anticipated processing delays. In doing so, it effectively balances blocking and waiting costs, while internalizing externalities of congestion.

To assess the empirical performance of the policy, we conduct extensive simulation experiments across a wide range of scenarios, including varying patterns of demand surges, uncertainty and errors in predictive information, and constraints on decision-making frequency. Across all settings, the proposed look-ahead policy consistently outperforms standard reactive benchmarks, yielding lower total costs and demonstrating strong robustness across a range of modeling assumptions and operational settings. We also perform a case study using real hospital data, which includes developing predictive models for ED arrivals and inpatient admission delays. The simulation results show that the look-ahead policy, when integrated with these predictive models, more effectively balances the trade-off between patient waiting and transfer rejection compared with commonly used reactive benchmark policies. This provides compelling evidence of the practical value of incorporating predictive information into ED admission control.

Taken together, our work makes two primary contributions. First, we rigorously derive and characterize a near-optimal admission control policy for transfer patients. Using a fluid approximation, we explicitly characterize how predictive information on future arrivals and inpatient admission delays enters the optimal decision rule through a hitting-time–based congestion metric. While the resulting policy exhibits a threshold structure consistent with classical admission control insights, our analysis precisely identifies the underlying congestion externality, establishes conditions for optimality, and provides a formal asymptotic optimality guarantee when the policy is implemented in the underlying stochastic system. This guarantee distinguishes our approach from purely simulation-based or *ad hoc* predictive admission rules and offers a principled foundation for how forecast information should be operationally incorporated into admission decisions. Second, we conduct a comprehensive empirical evaluation through extensive simulation experiments and

a high-fidelity case study using real hospital data. The results show that, when integrated with practical prediction models, the proposed look-ahead policy consistently outperforms commonly used reactive admission rules, achieving a more favorable trade-off between patient waiting and transfer rejection across a wide range of operating conditions. These findings provide empirical support for the operational benefits of systematically incorporating predictive information into ED admission control.

1.1. Literature reviews

Our work is mainly related to two lines of literature.

Admission control in queues Admission control has long been studied in queueing theory as a mechanism for mitigating congestion by balancing waiting costs against blocking or rejection penalties (see, for example, Heyman (1968), Lewis (2001), Altman et al. (2002), Ormeci (2004), Koçağa and Ward (2010)). In many classical settings, the optimal admission policy exhibits a simple threshold structure (Naor 1969, Stidham Jr 2002, Ward and Kumar 2008). Our work extends this literature by examining admission control in a queueing model tailored to the operational characteristics of an emergency department. In particular, our model incorporates time-varying arrival rates and admission delays arising from downstream resource constraints, resulting in a structure akin to a tandem system. Admission control for tandem systems has been studied in Silva et al. (2013), but their analysis focuses on a time-homogeneous tandem loss model.

Zayas-Cabán and Lewis (2020) study admission control in loss systems with periodically varying parameters, showing that optimal admission policies naturally tighten or relax in response to predictable demand cycles and quantifying the performance benefits of leveraging such periodic information. In another related direction, Bassamboo et al. (2005) analyze admission control in a multiclass many-server system with doubly stochastic arrivals. Using a multi-scale fluid limit in heavy traffic, they demonstrate that even under unpredictable rate fluctuations, effective admission decisions can be obtained by continually solving a simplified deterministic control problem informed by the most recent arrival-rate estimates. Similar to Ward and Kumar (2008) and Bassamboo et al. (2005), our analysis leverages a limiting approximation to obtain tractable insights. Since we study a transient system under pronounced time variation, a fluid model is particularly suitable and provides analytical tractability in our setting.

Our work also relates to a growing body of research on incorporating future information into admission or diversion decisions. Spencer et al. (2014) consider proactive redirectio in an $M/M/1$ queue and propose a look-ahead strategy using future-arrival information. Peng (2024) study admission control with arrival forecasts, characterizing how optimal policies change when predictive

information is available. In a queueing model motivated by ED operations, Xu and Chan (2016) demonstrate that look-ahead information can substantially improve diversion decisions. Our study differs from these works in several respects. We study admission control for inter-facility transfer requests, and explicitly model inpatient admission delays (ED boarding), along with other operational features that capture realistic ED dynamics. Importantly, beyond forecasting future demand, our framework also incorporates predictions of inpatient admission delays—a critical but often overlooked driver of congestion in ED settings. Relatedly, Bertsimas et al. (2022) operationalize machine-learning forecasts by embedding them into hospital EHR systems and deploying prediction-informed dashboards to support daily bed placement decisions. See also (Ata and Peng 2020, Delana et al. 2021, Ao et al. 2024) for additional studies that utilize future information for resource allocation and scheduling in stochastic systems.

Managing Congestion in the ED The conceptual model of Asplin et al. (2003) frames ED crowding in three interdependent components: input, throughput, and output. Crowding is essentially an imbalance where input exceeds throughput and output capacity in the acute care system. Extensive work in the emergency medicine literature has documented the causes and consequences of ED crowding, including its effects on patient safety, service delays, and provider well-being (see, for example, Hoot and Aronsky 2008, Bernstein et al. 2009, Morley et al. 2018). Practices in ED-to-ED transfers and acceptance decisions have also been studied (Greenwood-Ericksen et al. 2025).

Within operations research, ED congestion has been studied from multiple angles using stochastic modeling, empirical analysis, and simulation. Queueing models have been widely applied to support operational decisions. For example, Green et al. (2006) and Yom-Tov and Mandelbaum (2014) use queueing analysis to determine appropriate staffing levels under varying demand conditions. Jacobson et al. (2012) and Saghafian et al. (2012) show that streaming and prioritization schemes can be effective tools for improving system performance, particularly when the ED experiences extreme levels of crowding. Shi et al. (2016) develop a stochastic network model of inpatient operations and show that inpatient discharge policy can have a substantial impact on ED boarding times. A parallel empirical literature investigates how ED clinicians adapt their behavior in response to congestion. For example, Song et al. (2015) analyze how alternative queue configurations influence ED physician productivity. Batt and Terwiesch (2017) document the phenomenon of early task initiation, where upstream staff (e.g., triage nurses) begin tasks typically performed by downstream providers when the ED becomes busy. Ding et al. (2019) and Li et al. (2021) show that prioritization decisions also shift with crowding. Adepoju et al. (2023) studied how reactive hospital-wide responses to ED crowding may have unintended negative consequences. Recently, Canellas et al. (2025) consider fair patient prioritization and placement in ED operations, using

a predictive–prescriptive MILP that breaks predicted ED length of stay into actionable components. Our work contributes to this literature by examining a complementary and understudied operational lever: admission control for inter-facility transfer patients.

1.2. Notations

For integer $n \geq 1$, we denote $[n] = \{1, 2, \ldots, n\}$ as the set of integers from 1 to n. For $x \in \mathbb{R}$, denote $x_+ = \max\{x, 0\}$. For set S, denote $|S|$ as its cardinality.

2. Problem Formulation

We model the emergency department (ED) as a multi-server queue with a potentially time-varying number of servers, denoted by $c(t)$. Patient arrivals consist of two distinct streams: (1) External Arrivals, which include patients entering the ED from outside the hospital system, such as walk-ins, ambulance arrivals, and referrals from external healthcare providers; and (2) Transfer Requests, which consist of patients transferred from other EDs. External arrivals follow a time-varying Poisson process with rate $\lambda_1(t)$, whereas transfer requests arrive according to a Poisson process with rate $\lambda_2(t)$. We write $\lambda(t) = \lambda_1(t) + \lambda_2(t)$ as the aggregated arrival rate. Patients are served on a first-come-first-served basis, with service times that are exponentially distributed with rate μ. While waiting for service, patients may abandon the queue if their patience expires; patience times are assumed to be exponentially distributed with rate $\theta < \mu$. Upon completing ED treatment, patients either leave the hospital directly with probability p or are admitted to the inpatient wards (IWs) with probability $1 - p$. Patients requiring inpatient admission experience a boarding delay prior to transfer, reflecting limited bed availability or administrative bottlenecks on the inpatient side. These boarding times are modeled as exponentially distributed with a time-varying rate $\nu(t)$, which we refer to as the processing rate. During this delay, boarding patients continue to occupy ED servers.

We consider an admission control policy under which the hospital determines whether to accept incoming transfer requests. Specifically, the control variable $G(t) \in \{0, 1\}$ indicates whether transfer requests arriving at time t are accepted, $G(t) = 1$, or rejected, $G(t) = 0$.

Let $X(t)$ denote the total number of patients in the ED who have not yet completed service (treatment), including both those waiting in the queue and those currently receiving service. In addition, let $Z(t)$ denote the number of patients currently in service and $Q(t)$ denote the number of patients waiting in the queue. Then, $X(t) = Q(t) + Z(t)$. We further let $B(t)$ denote the number of patients boarding in the ED while awaiting transfer to the IWs.

Let A_1, A_2, S, D, and H be independent rate-1 Poisson processes representing external arrivals, transfer requests, service completions, patient abandonments, and inpatient transfer completions, respectively. The system dynamics can then be characterized by

$$\begin{aligned} X(t) =& X(0) + A_1\left(\int_0^t \lambda_1(t)dt\right) + A_2\left(\int_0^t \lambda_2(t)G(t)dt\right) - S\left(\int_0^t \mu Z(t)dt\right) - D\left(\int_0^t \theta Q(t)dt\right), \\ B(t) =& B(0) + S\left(\int_0^t (1-p)\mu Z(t)dt\right) - H\left(\int_0^t \nu(t)B(t)dt\right), \end{aligned} \tag{1}$$

Without loss of optimality, we assume the system is work-conserving, so that

$$Z(t) = \min\{c(t) - B(t),\, X(t)\}.$$

An admission control policy $G(t)$ may depend on the current system state $(X(t), B(t))$, the future arrival and processing rates $\{\lambda(s) : s \geq t\}$ and $\{\nu(s), s \geq t\}$ (or their forecasts), and future staffing levels $\{c(s), s \geq t\}$, reflecting the fact that staffing decisions are typically made in advance. Let π denote an admissible admission control policy, and use superscripts to indicate dependence on π, e.g., $X^\pi(t)$, $B^\pi(t)$.

We associate a holding cost h per patient per unit time spent waiting in the queue, and a blocking cost l for each rejected transfer request. The objective is to minimize the expected total cost over a planning horizon T:

$$V^\pi(x,b) = \mathbb{E}\left[\int_0^T h(X^\pi(t) - Z^\pi(t))dt + lL^\pi(T) \mid X(0) = x, B(0) = b\right] \tag{2}$$

where $L^\pi(T)$ denotes the number of transfer requests rejected by time T. Note that

$$L^\pi(T) \sim \tilde{E}\left(\int_0^T \lambda_2(t)\left(1 - G^\pi(t)\right)dt\right) \tag{3}$$

where $\tilde{E}$ is a rate-1 Poisson process. The planning horizon T is chosen to be sufficiently long for the system to fully recover from any demand surge; a formal definition is provided in Appendix A.

The model deliberately abstracts from certain operational details of patient flow, such as unit-level bed heterogeneity, service prioritization, individual patient-level acuity, and non-exponential distributions. Our objective is not to replicate the full complexity of ED patient flow, but to isolate the central tradeoff that governs transfer admission decisions: accepting transfer patients increases future congestion, whereas rejecting them incurs an immediate blocking cost. This tradeoff operates at an aggregate level and is driven by the interaction among current system congestion, anticipated future demand, and expected boarding times. From an operational standpoint, admission control decisions for transfer patients are typically made at a relatively coarse temporal and organizational

scale, based on aggregate congestion indicators, such as ED census, boarding levels, and expected discharges, rather than detailed unit-level state information. The state variables tracked in our model, namely the total ED census and the boarding population, are therefore well aligned with the information routinely available to decision makers and the level at which such decisions are implemented in practice.

3. Fluid Analysis

The stochastic control problem in (2) is analytically intractable due to (i) the nonstationary primitives $(\lambda_1(t), \lambda_2(t), \nu(t), c(t))$ and (ii) the complexity induced by the joint evolution of the ED census and boarding population. Moreover, even when numerical solutions are feasible for stylized instances, they offer limited structural insight into the role of predictive information in admission decisions.

To obtain a tractable characterization of the optimal policy and to clarify how future information enters admission decisions, we study a deterministic fluid approximation. The fluid model captures the mean dynamics of the underlying queueing system and is particularly appropriate in the congested, time-varying regimes of interest.

Let $x(t)$, $z(t)$, $b(t)$, and $g(t)$ denote the fluid analogues of $X(t)$, $Z(t)$, $B(t)$, and $G(t)$, respectively. The fluid dynamics satisfy

$$\begin{aligned} \dot{x}(t) &= \lambda_1(t) + g(t)\lambda_2(t) - \mu z(t) - \theta\big(x(t) - z(t)\big), \\ \dot{b}(t) &= (1-p)\mu z(t) - \nu(t) b(t), \end{aligned} \tag{4}$$

where $z(t) = \min\{c(t) - b(t), x(t)\}$ and $\dot{x}(t) = dx(t)/dt$, $\dot{b}(t) = db(t)/dt$.

3.1. Single Congestion Shock

We first analyze a setting with a single congestion shock beginning at time 0 and ending at time $\kappa > 0$. Such a shock may be driven by a demand surge, a temporary reduction in the processing rate, or both. We impose the following regularity and load conditions.

ASSUMPTION 1. *1. The functions $\lambda_1(t)$, $\lambda_2(t)$, $\nu(t)$, and $c(t)$ are piecewise monotone with a finite number of pieces.*

2. The initial condition satisfies $x(0) + b(0) \geq c(0)$. There exists $\epsilon_1 > 0$ such that for all $t \in [0, \kappa)$,

$$\lambda_1(t) - \max\{p\mu c(t), \nu(t)c(t)\} \geq \epsilon_1.$$

3. There exists $\epsilon_2 > 0$ such that for all $t \geq \kappa$,

$$\lambda(t) - \min\{p\mu c(t), \nu(t)c(t)\} \leq -\epsilon_2,$$

where $\lambda(t) = \lambda_1(t) + \lambda_2(t)$.

Assumption 1.1 is a standard regularity condition ensuring existence and uniqueness of solutions to the fluid ODEs and enabling convergence arguments under many-server scaling. Assumptions 1.2–3 impose a pronounced overload prior to time κ and a subcritical regime thereafter, which together ensure that the fluid queue remains positive before κ and is eventually cleared after κ. While these conditions can be relaxed, the essential requirement for our analysis is that the fluid queue does not hit zero during the overload phase and eventually drains to zero in finite time.

We assume the decision maker has access to the functions $\{\lambda_1(s), \lambda_2(s), \nu(s), c(s) : s \geq t\}$ (or forecasts thereof) when choosing $g(t)$. This is *not* an assumption of knowing the realized future sample path of the stochastic primitives in (1); rather, it corresponds to having access to their conditional expectations, as produced by a calibrated forecasting model. Staffing levels $c(t)$ are assumed known in advance.

Let $q(t) = x(t) - z(t)$ denote the fluid queue length. Define the (fluid) queue-clearing time

$$\tau \;=\; \inf\{t \geq \kappa :\; x(t) + b(t) = c(t)\}.$$

Under Assumption 1, $\tau < \infty$ and $q(t) = 0$ for all $t \geq \tau$. The fluid optimal control problem corresponding to (2) can be written as

$$\begin{aligned} \min_{g(\cdot)} \quad & \int_0^\tau \Big[h\big(x(t) - z(t)\big) + l\lambda_2(t)\big(1 - g(t)\big)\Big]\, dt \\ \text{s.t.} \quad & (4), \qquad z(t) = \min\{x(t), c(t) - b(t)\}, \qquad 0 \leq g(t) \leq 1. \end{aligned} \tag{5}$$

For $t \geq 0$, define the hitting time $H_t(x, b)$ as the time required for the system to return to a zero-queue state when *all* transfers are accepted (i.e., $g(\cdot) \equiv 1$), starting from state (x, b) at time t:

$$H_t(x, b) := \inf\Big\{\Delta > 0 :\; \tilde{x}(t + \Delta) + \tilde{b}(t + \Delta) \leq c(t + \Delta) \;\Big|\; \tilde{x}(t) = x,\; \tilde{b}(t) = b\Big\}, \tag{6}$$

where $(\tilde{x}(\cdot), \tilde{b}(\cdot))$ is the solution to the uncontrolled ODE system

$$\begin{aligned} & \tilde{x}(t) = x, \qquad \tilde{b}(t) = b, \\ & \dot{\tilde{x}}(u) = \lambda(u) - \mu\big(c(u) - \tilde{b}(u)\big) - \theta\, \tilde{q}(u), \qquad u \geq t, \\ & \dot{\tilde{b}}(u) = (1 - p)\mu\big(c(u) - \tilde{b}(u)\big) - \nu(u)\tilde{b}(u), \qquad u \geq t, \end{aligned} \tag{7}$$

with $\tilde{q}(u) = \tilde{x}(u) - \min\{\tilde{x}(u), c(u) - \tilde{b}(u)\}$. For boundary states satisfying $x + b = c(t)$, we define $H_t(x,b)$ by continuity as $H_t(x,b) = \lim_{x \downarrow (c(t)-b)} H_t(x,b)$. Intuitively, $H_t(x,b)$ measures the duration of congestion implied by the current state and the future primitives $(\lambda(\cdot), \nu(\cdot), c(\cdot))$.

The following theorem characterizes the optimal admission control for (5).

THEOREM 1. *Under Assumption 1, an optimal control for (5) is given by*

$$g^*(t) = \mathbf{1}\left\{\frac{h}{\theta}\Big(1 - \exp(-\theta H_t(x(t), b(t)))\Big) \le l\right\}. \tag{8}$$

Theorem 1 implies that transfer requests should be rejected whenever the congestion-adjusted holding cost, $\frac{h}{\theta}\Big(1 - e^{-\theta H_t(x(t),b(t))}\Big)$, exceeds the blocking cost l. Note that if $h/\theta \le l$, the blocking cost is very large compared to the holding cost if everyone waits until abandonment. In this case, we will never reject transfer patients. In the special case $\theta = 0$, taking the limit $\theta \downarrow 0$ yields the simpler condition $h\, H_t(x(t), b(t)) \le l$. These comparisons emphasize that the relevant holding cost is not the instantaneous queue length alone, but rather the expected congestion duration induced by current conditions, future demand, and boarding delays.

The proof of Theorem 1 is given in Appendix B.1, where we apply Pontryagin's Minimum Principle and verify the sufficient conditions for optimality.

3.2. Multiple Congestion Shocks

The preceding analysis extends to settings with multiple congestion shocks. For clarity, we present the case of two shocks; extensions to more than two shocks follow analogously. Let

$$0 < \kappa_a < \kappa_b < \kappa_c,$$

where the system is overloaded on $[0, \kappa_a)$ and $[\kappa_b, \kappa_c)$ and underloaded on $[\kappa_a, \kappa_b)$ and $[\kappa_c, \infty)$. We impose the following conditions.

ASSUMPTION 2. *1. The functions $\lambda_1(t)$, $\lambda_2(t)$, $\nu(t)$, and $c(t)$ are piecewise monotone with a finite number of pieces.*

2. The initial condition satisfies $x(0) + b(0) \ge c(0)$. There exists $\epsilon_1 > 0$ such that for all $t \in [0, \kappa_a) \cup [\kappa_b, \kappa_c)$,

$$\lambda_1(t) - \max\{p\mu c(t),\, \nu(t)c(t)\} \ge \epsilon_1.$$

3. There exists $\epsilon_2 > 0$ such that for all $t \in [\kappa_a, \kappa_b) \cup [\kappa_c, \infty)$,

$$\lambda(t) - \min\{p\mu c(t),\, \nu(t)c(t)\} \le -\epsilon_2.$$

Define the ultimate clearing time $\tau = \inf\{t > \kappa_c : \ x(t) + b(t) = c(t)\}$. The next theorem shows that the optimal policy retains the same functional form as in the single-shock case, with the same hitting-time $H_t(\cdot)$ defined in (6).

THEOREM 2. *Under Assumption 2, an optimal control for (5) is given by*

$$g^*(t) = \mathbf{1}\left\{\frac{h}{\theta}\Big(1 - \exp(-\theta H_t(x(t), b(t)))\Big) \le l\right\}.$$

The proof of Theorem 2 is provided in Appendix B.2.

The policy structure derived above is driven by the ability to quantify the anticipated duration of congestion through a hitting-time metric. While our analysis is conducted under the specific parametric assumptions of (4), the underlying logic of the policy does not hinge on these assumptions *per se*, but rather on the existence of a deterministic approximation that links current congestion and future system primitives to the time required for the system to clear. In more complicated settings, e.g., with multiple patient classes and heterogeneous service dynamics, one can formulate the corresponding fluid model and numerically compute an analogous hitting-time metric. While we do not derive formal analytical results for these extensions, this perspective suggests that the same threshold-based admission logic may continue to apply, with increased modeling complexity primarily affecting the computation of $H_t(\cdot)$ rather than the qualitative structure of the policy. We explore this idea further through a case study in Section 6, illustrating how the proposed framework can be adapted beyond the baseline model.

4. Admission Control Policies for the Stochastic System

In this section, we describe how the fluid-optimal policy can be implemented in the original stochastic system and discuss several practical considerations that arise in operational settings, including prediction error and limited decision-update frequency.

4.1. Fluid-Inspired Admission Control Rule

The fluid analysis yields a simple, interpretable admission rule that can be directly applied to the stochastic system. At any time t, the decision maker evaluates the condition

$$\frac{h}{\theta}\left(1 - \exp\Big(-\theta H_t\big(X(t), B(t)\big)\Big)\right) > l.$$

If the inequality holds, incoming transfer requests are rejected; otherwise, they are accepted. Here, $(X(t), B(t))$ denote the current stochastic system state, and $H_t(\cdot)$ is the congestion hitting-time metric computed from the fluid model. Operationally, $H_t(X(t), B(t))$ represents the predicted time required for congestion to clear if all transfer requests were accepted from time t onward.

This policy has several appealing features from an implementation standpoint. First, it depends only on low-dimensional, routinely monitored state variables (ED census and boarding population). Second, future arrival rates, processing rates, and staffing levels enter the decision only through the scalar quantity H_t, providing a natural way to aggregate predictive information into a single operational signal.

Although the admission control rule is derived from a deterministic fluid approximation, it admits a rigorous justification for the underlying stochastic system. In particular, under a many-server scaling regime in which arrival rates and capacity grow proportionally, the fluid-translated policy is asymptotically optimal. This result provides formal theoretical support for using the fluid hitting time as the decision-relevant quantity for admission control in large, highly-utilized EDs. The precise statement of the asymptotic optimality result, along with the technical assumptions and proof, is provided in Appendix A.

4.2. Prediction Error and Imperfect Information

In practice, future arrival rates $\lambda(t)$ and processing rates $\nu(t)$ are not known exactly and must be estimated using predictive models. Let $\hat{\lambda}(t)$ and $\hat{\nu}(t)$ denote such predictions. When forecasts are available, we compute an approximate hitting time $\widehat{H}_t$ by substituting $\hat{\lambda}(t)$ and $\hat{\nu}(t)$ into the fluid dynamics, and apply the same admission rule:

$$\frac{h}{\theta}\left(1-\exp\left(-\theta\widehat{H}_t\big(X(t),B(t)\big)\right)\right) > l.$$

When reliable short-term forecasts are unavailable, a simple alternative is to approximate future arrival and processing rates using coarse summaries of the anticipated operating environment. There are many reasonable ways to construct such approximations. As one illustrative choice, we consider the average future rates over the remaining planning horizon,

$$\bar{\lambda} = \frac{1}{T-\kappa}\int_{\kappa}^{T}\lambda(t)\,dt, \qquad \bar{\nu} = \frac{1}{T-\kappa}\int_{\kappa}^{T}\nu(t)\,dt,$$

and compute an approximate hitting time $\bar{H}_t$ by evaluating the fluid dynamics under constant rates $\lambda(s)=\bar{\lambda}$ and $\nu(s)=\bar{\nu}$ for $s \geq t$. This approximation yields a simple, state-dependent admission rule that reflects the average congestion-clearing capacity expected over the remainder of the planning horizon.

Several features of the policy help mitigate the impact of prediction error. First, the current system state $(X(t), B(t))$ already encodes real-time congestion information, which partially offsets inaccuracies in forecasts of future arrival and processing rates. Second, the hitting-time metric

aggregates these rates over a future horizon, causing small, idiosyncratic prediction errors to average out. Finally, because decisions depend only on whether H_t crosses a fixed threshold, moderate prediction errors typically do not alter the admission decision when the system is far from the switching boundary. When H_t lies close to the threshold, prediction errors may affect the decision; however, in this regime the marginal congestion cost and blocking cost are nearly balanced, so the performance difference between accepting and rejecting transfer requests is small. Overall, the policy is relatively robust to modest prediction error, though large or systematic biases can still degrade performance. We investigate this sensitivity through numerical experiments in Section 5.2.

4.3. Periodic Review and Implementation Considerations

In operational environments, admission decisions are rarely updated continuously. Moreover, frequent re-evaluation can lead to undesirable decision volatility driven by stochastic fluctuations in the system state, causing the hitting-time metric H_t to oscillate around the decision threshold. To address these practical considerations, we consider a periodic review implementation in which admission decisions are updated at discrete intervals of length δ.

Specifically, at review times $k\delta$, $k = 0, 1, 2, \ldots$, the admission decision is fixed over the interval $[k\delta, (k+1)\delta)$. Transfer requests are rejected during this interval if

$$H_t\big(X(k\delta), B(k\delta)\big) - a\delta > -\frac{1}{\theta}\log\left(1 - \frac{l\theta}{h}\right), \tag{9}$$

where $a \geq 0$ is a tuning parameter introduced to compensate for the time lag between reviews.

The review interval δ governs a trade-off between responsiveness and stability. Smaller values of δ allow the policy to react quickly to changes in congestion, but may result in frequent switching driven by noise. Larger values of δ yield more stable decisions and lower administrative burden. Importantly, because the policy depends on whether H_t exceeds a fixed threshold, decision quality is relatively insensitive to δ when H_t is well above or well below the threshold. When H_t is close to the threshold, the performance difference between accepting and rejecting transfers is small, so less frequent updates can help reduce volatility without materially affecting outcomes. We illustrate these effects through numerical experiments in Sections 5 and 6.

5. Numerical Experiments

In this section, we evaluate the performance of the proposed look-ahead admission control policy through a set of synthetic numerical experiments. Our primary objective is to quantify the operational value of incorporating predictive information into transfer admission decisions, relative to

policies that rely solely on contemporaneous congestion measures. Because admission control inherently involves a trade-off between congestion mitigation and access for transfer patients, policy performance cannot be meaningfully evaluated at a single operating point. Accordingly, throughout this section we summarize performance using trade-off curves between the average waiting time and the fraction of rejected transfer requests.

Policy comparison. We compare the proposed look-ahead policy against a class of state-dependent benchmark policies motivated by current hospital practice and by optimal admission control rules in queueing models without predictive information. These benchmark policies rely exclusively on contemporaneous congestion indicators and do not incorporate forecasts of future demand or processing capacity.

Under a benchmark policy, an incoming transfer request is rejected whenever

$$\eta X(t) + B(t) \le \Gamma,$$

where $X(t)$ denotes the total ED census, $B(t)$ the boarding population, and η and Γ are tunable parameters. This class captures a wide range of commonly used threshold rules based on ED occupancy and boarding levels.

For any fixed η, varying Γ traces out the achievable trade-off between congestion (measured by the average waiting time) and access (measured by transfer rejection rate) within this class of state-based policies. The resulting trade-off curve summarizes the attainable performance of policies that rely only on contemporaneous system information and do not use predictive inputs, providing a natural and fair benchmark for assessing the incremental value of prediction.

We evaluate benchmark policies for $\eta \in \{0, 0.5, 1\}$ and $\Gamma \in \{10, 11, \ldots, 100\}$. Once Γ is appropriately tuned, benchmark performance is largely insensitive to the choice of η (see Appendix D for details). For clarity, we therefore present results for $\eta = 0$.

For the proposed look-ahead policy, we construct an analogous trade-off curve by varying the ratio l/h, which governs the relative penalty placed on transfer rejection versus patient waiting.

Experimental design and baseline parameters. Unless otherwise noted, we consider a system with constant service capacity $c(t) = 30$ over a horizon $T = 150$. A fraction $\gamma = 0.2$ of arrivals correspond to transfer requests, with the remaining arrivals representing non-transfer patients. Service times are exponential, with the service rate normalized to $\mu = 1$, so that time is measured in units of mean service time. The abandonment rate is $\theta = 0.1$, and patients are discharged directly from the ED with probability $p = 0.5$. The system is initialized at $(X(0), B(0)) = (30, 0)$. We normalize the waiting cost to $h = 1$ and vary the blocking cost $l \in [0, 10)$ to trace out the tradeoff curves. Admission

decisions are updated periodically, with review intervals $\delta \in \{1, 5\}$ and $a = 0.5$ to compensate for the time lag between reviews. All reported performance measures are estimated as averages over 200 independent sample paths.

In the main paper, we fix the inpatient processing rate at $\nu(t) = 0.5$ to isolate the effect of predictive demand information. Appendix D considers scenarios with time-varying and reduced processing rates; across these settings, the qualitative comparison between policies remains unchanged. We assume the processing rate is known without error, i.e., $\hat{\nu}(t) = \nu(t)$.

Arrival pattern and prediction error. We assess policy performance along two dimensions. First, we consider different arrival patterns corresponding to a single congestion surge and two congestion surges, which capture distinct temporal structures of demand shocks. Second, and more importantly, we examine how different forms of prediction error in arrival rate forecasts affect policy performance.

In all congestion-surge scenarios, admission decisions are based on predicted arrival rates rather than perfect future information. Predicted arrival rates are generated on a discrete time grid with step size $\Delta t = 0.1$ and held constant between updates, with nonnegativity enforced by truncation.

5.1. Congestion surge scenarios

We consider two demand environments that differ in the temporal structure of congestion surges: a single-surge scenario and a two-surge scenario. In both cases, we first assume unbiased, independently distributed prediction errors in the arrival rate forecasts:

$$\hat{\lambda}(k\Delta t) = \lambda(k\Delta t) + \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, \sigma^2).$$

We examine more general forms of prediction error in Section 5.2.

5.1.1. Single congestion shock We first consider a scenario with a single congestion shock, where the arrival rate $\lambda(t)$ evolves as

$$\lambda(t) = \begin{cases} 2, & 0 \le t \le 50, \\ 20, & 50 < t \le 100, \\ 2, & 100 < t \le T, \\ 0, & \text{otherwise.} \end{cases}$$

Figure 1 shows a representative realization of the predicted arrival rate $\hat{\lambda}(t)$.

Figure 2 compares the performance of the benchmark and look-ahead policies by plotting the trade-off between total patient waiting time and the proportion of rejected transfer requests. Across both panels, corresponding to review intervals $\delta = 1$ and $\delta = 5$ respectively, the look-ahead policy

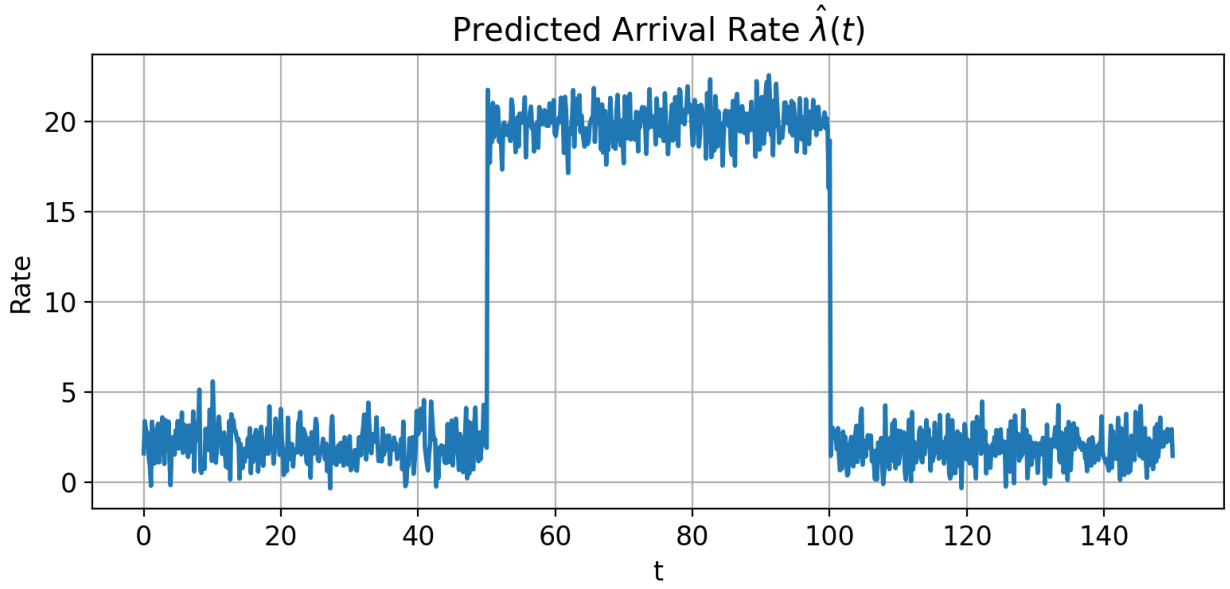

Figure 1 Single congestion shock: predicted arrival rate $\hat{\lambda}(t) = \lambda(t) + \epsilon(t)$, where $\epsilon(t) \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 1$.

uniformly dominates the benchmark. Specifically, for any given level of waiting time, the look-ahead policy achieves a lower rejection rate, and conversely, for any given rejection rate, it yields shorter waiting times. These results highlight the operational value of incorporating predictive demand information into dynamic admission decisions, particularly in environments subject to transient congestion shocks.

5.1.2. Two congestion shocks We next consider a setting with two congestion shocks, where the arrival rate $\lambda(t)$ evolves as

$$\lambda(t) = \begin{cases} 2, & 0 \le t \le 30, \\ 20, & 30 < t \le 60, \\ 2, & 60 < t \le 90, \\ 20, & 90 < t \le 120, \\ 2, & 120 < t \le T, \\ 0, & \text{otherwise.} \end{cases}$$

Figure 3 illustrates a representative realization of the corresponding predicted arrival rate $\hat{\lambda}(t)$.

Figure 4 compares the performance of the benchmark and look-ahead policies by plotting the trade-off between total patient waiting time and the fraction of rejected transfer requests. For both review intervals $\delta = 1$ and $\delta = 5$, the look-ahead policy again consistently dominates the benchmark. This demonstrates that the benefits of predictive admission control persist even in more complex environments with multiple congestion surges and intervening recovery periods.

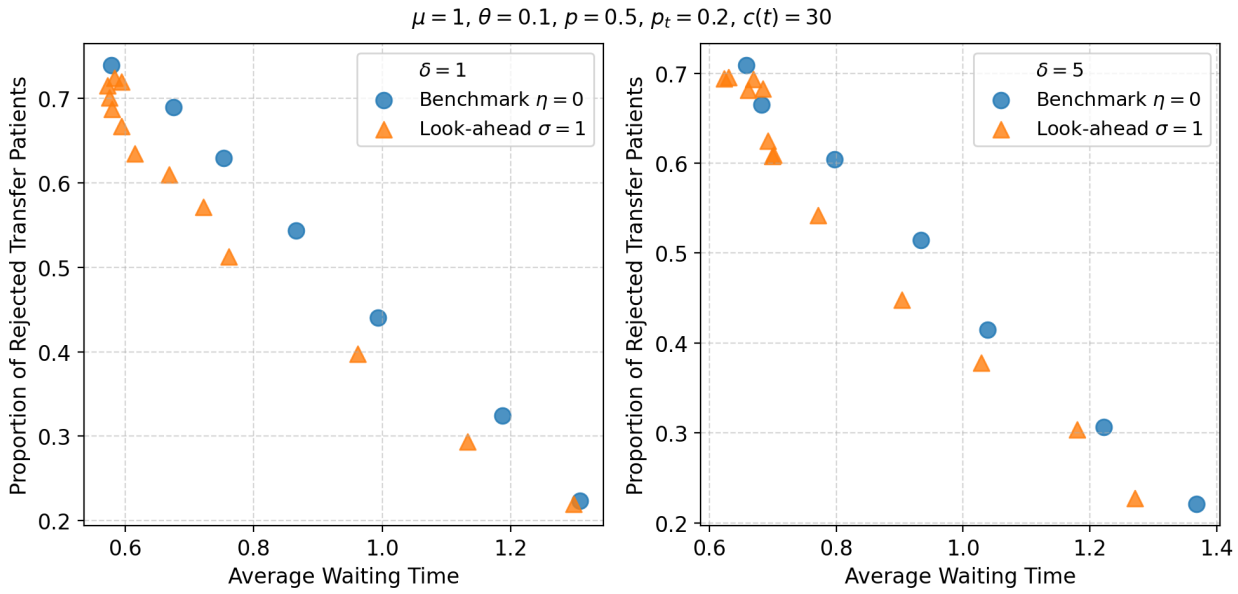

Figure 2 Single congestion shock: comparison between look-ahead v.s. benchmark.

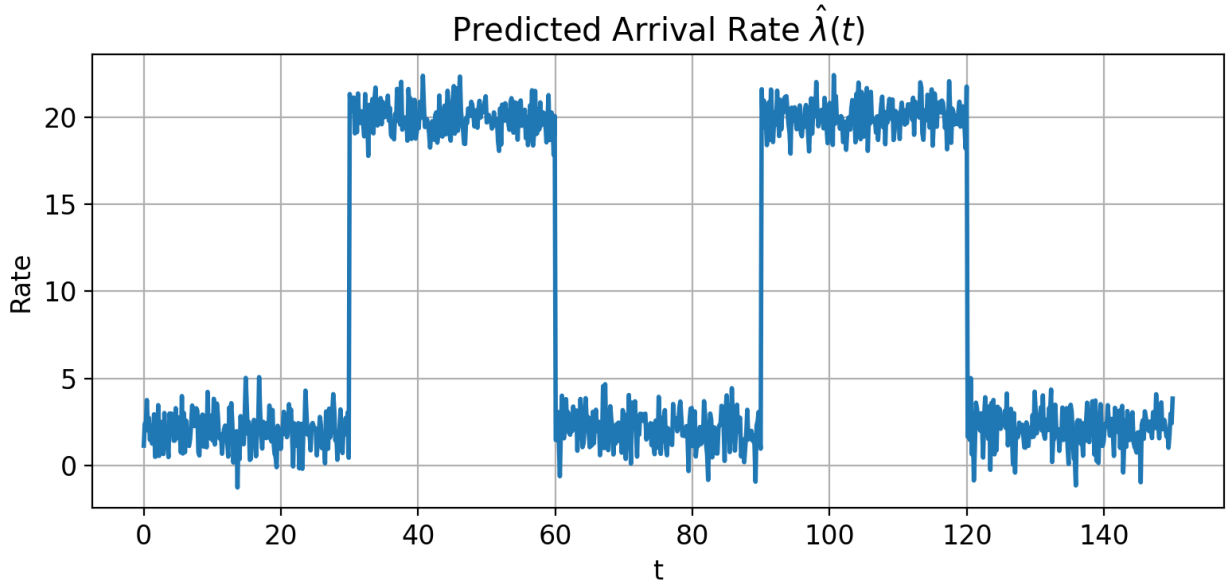

Figure 3 Two congestion shocks: predicted arrival rate $\hat{\lambda}(t) = \lambda(t) + \epsilon(t)$, where $\epsilon(t) \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 1$.

5.2. Prediction Errors

In Section 5.1, we evaluated policy performance under a baseline prediction error model with unbiased, independently distributed noise in the arrival rate forecasts. We now extend this analysis to examine robustness under a broader range of forecast errors, including serial correlation, horizon-dependent uncertainty, systematic bias, and misprediction of surge timing.

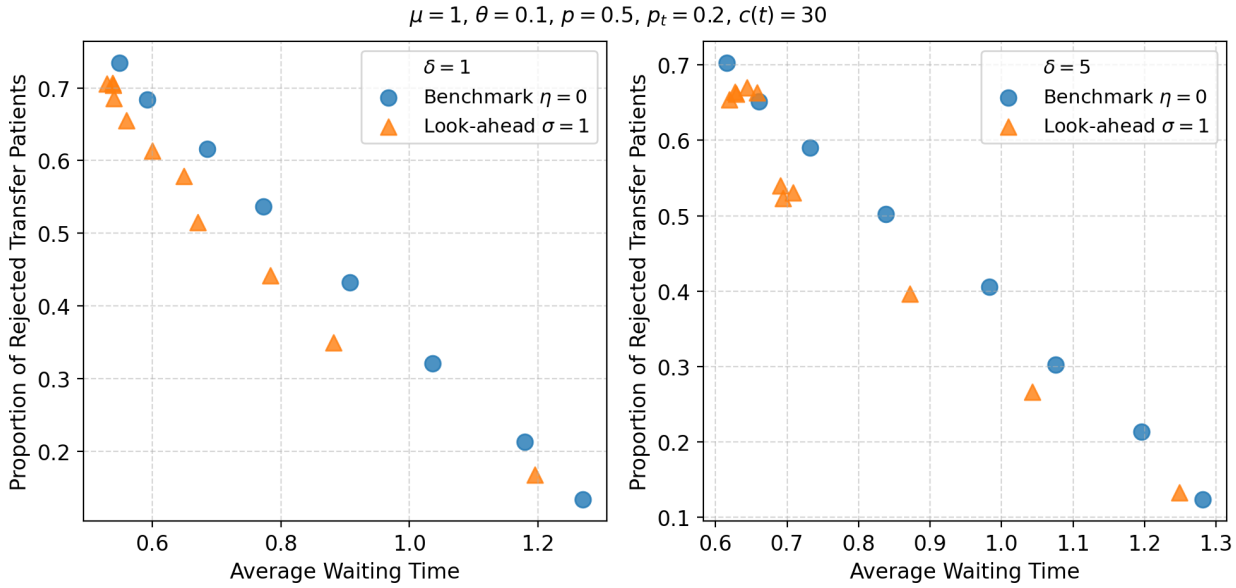

Figure 4 Two congestion shocks: comparison between look-ahead v.s. benchmark.

Scenario I: Independent forecast errors (baseline). For each update step,

$$\hat{\lambda}(k\Delta t) = \lambda(k\Delta t) + \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, \sigma^2).$$

This specification captures unbiased but noisy forecasts updated periodically and serves as the baseline benchmark. We consider $\sigma \in \{0, 1, 2, 4\}$ to assess sensitivity to the noise level.

Scenario II: Correlated forecast errors. For each update step,

$$\hat{\lambda}(k\Delta t) = \lambda(k\Delta t) + \epsilon_k, \quad \epsilon_k = \rho\epsilon_{k-1} + u_k, \quad u_k \sim \mathcal{N}(0, \sigma_u^2),$$

where $\rho \in \{0.7, 0.99\}$ and $\sigma_u^2 = (1-\rho^2)\sigma^2$.

Scenario III: Horizon-dependent forecast uncertainty. For each update step,

$$\hat{\lambda}(k\Delta t) = \lambda(k\Delta t) + \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, \sigma^2(1+\beta k)),$$

with $\beta \in \{0.02, 0.2\}$.

Scenario IV: Biased forecasts. For each update step,

$$\hat{\lambda}(k\Delta t) = (1+\alpha)\lambda(k\Delta t) + \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, \sigma^2),$$

with $\alpha \in \{-0.6, -0.4, -0.2, 0.2, 0.4, 0.6\}$. Positive values of α correspond to persistent overestimation of demand, while negative values represent underestimation.

Scenario V: Shock timing misprediction. For each update step,

$$\hat{\lambda}(k\Delta t) = \lambda(k\Delta t - \xi) + \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, \sigma^2),$$

with $\xi \in \{5, 10\}$.

For Scenarios II–V, we consider $\sigma \in \{1, 2\}$. For ease of exposition, the main paper reports results for the two-surge scenario with $\sigma = 2$ and review interval $\delta = 1$. Results for additional cases, including the single-surge scenario and alternative σ and δ, are qualitatively similar and reported in Appendix D.

Across the baseline model (Scenario I) and its extensions in Scenarios II, III, and IV with positive α, policy performance remains qualitatively stable. As shown in Figure 5, we do not observe systematic degradation in the trade-off between waiting time and transfer rejection. This indicates robustness to moderate levels of unbiased noise, correlated errors, increasing forecast uncertainty, and conservative bias that preserves information about surge magnitude.

In contrast, performance deteriorates under systematic underestimation of demand (Scenario IV with negative α) and under shock timing misprediction (Scenario V), as illustrated in Figure 6. Because forecasts are constrained to be nonnegative, negative bias effectively attenuates the perceived severity of congestion, while timing errors delay the policy response to surges. These forms of misspecification reduce the effectiveness of anticipatory admission control and highlight the importance of accurately capturing both the magnitude and timing of congestion events.

6. Case Study: Implementation with Real Hospital Data

In this section, we evaluate the proposed look-ahead admission control policy using operational data from a large academic medical center. The purpose of this case study is to assess whether the policy can be implemented with realistically available information, including data-driven forecasts of demand and boarding delays, and whether it yields measurable improvements relative to the hospital's current transfer-admission practice when calibrated to empirical operating conditions.

A central practical consideration is that transfer patients differ systematically from non-transfer patients in their likelihood of requiring inpatient admission. To capture this empirically important heterogeneity while preserving the policy logic developed earlier, Section 6.1 generalizes the base model to allow type-dependent admission probabilities. This extension modifies only the fluid dynamics used to compute the hitting time; the admission decision remains a threshold rule in the predicted hitting time.

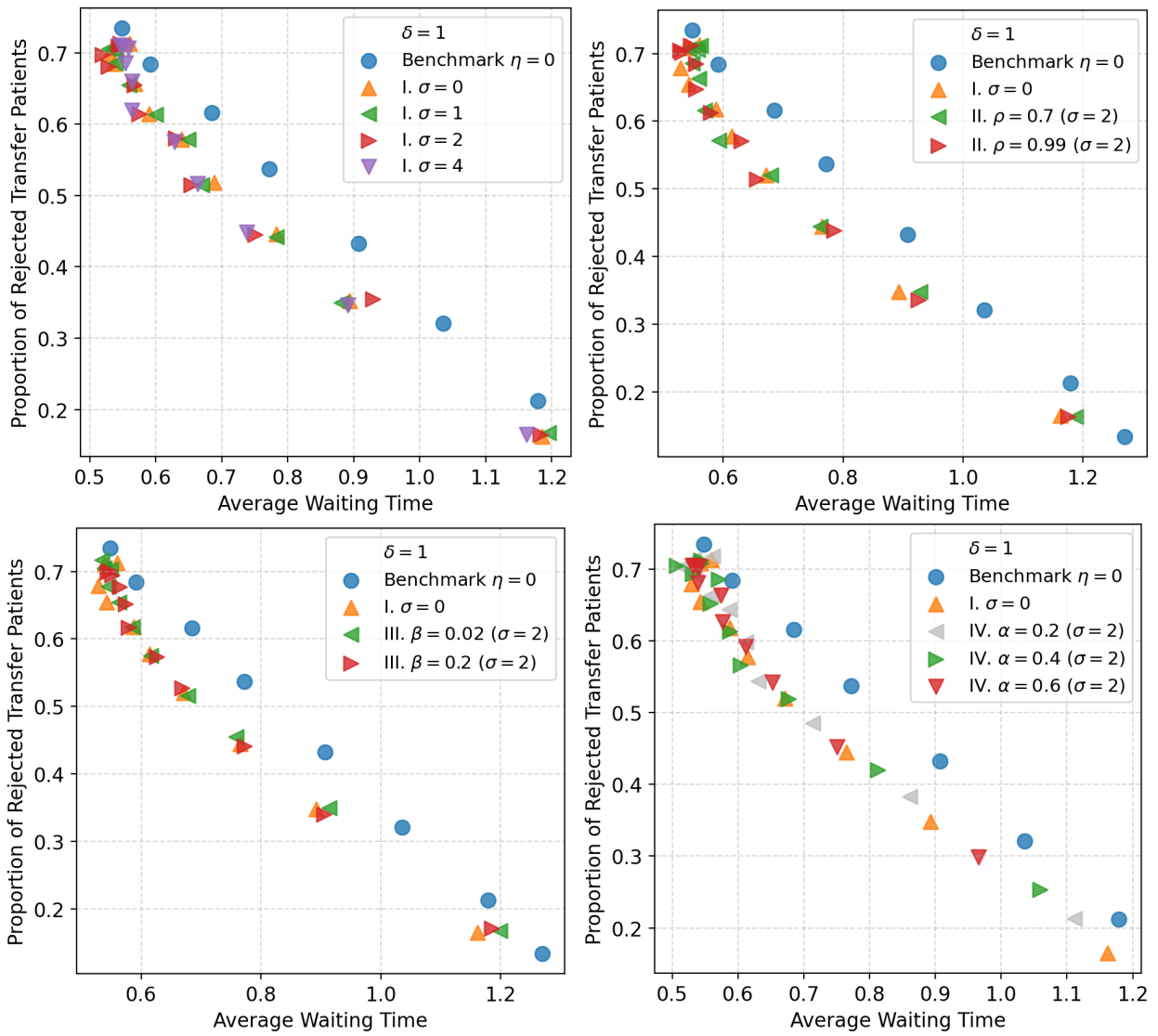

Figure 5 Similar performance with prediction errors in scenarios I, II, III, IV with positive α

Building on this implementation model, Section 6.2 describes the data sources, forecasting models for arrivals and processing rates, and parameter calibration. Section 6.3 then compares the performance of the look-ahead policy to the hospital's current policy using simulated counterfactuals over the study period. Together, the results illustrate how predictive information can be operationalized through a simple periodic-review rule by quantifying the improvements in the congestion-access trade-off.

6.1. Operational Model with Heterogeneous Admission Outcomes

To support implementation, we generalize the base model by allowing the probability of inpatient admission to depend on patient type. Specifically, patients are classified into external patients (type 1) and inter-facility transfer patients (type 2). Let p_1 and p_2 denote the probabilities that

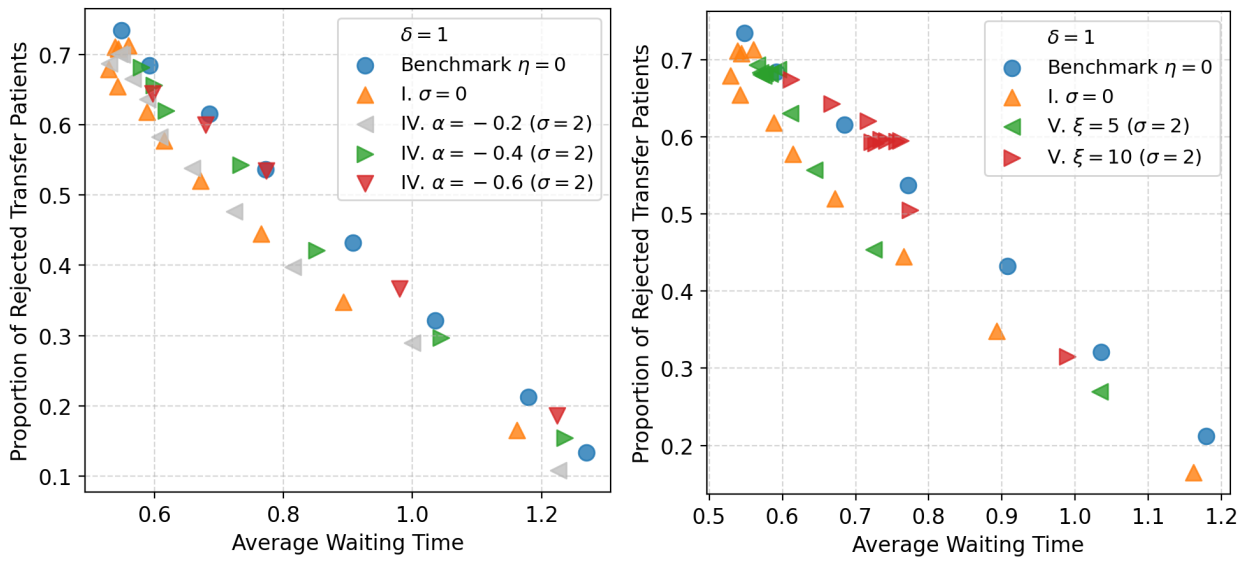

Figure 6 Degradation in performance with prediction errors in scenarios IV (with negative α) and V

type-1 and type-2 patients, respectively, are discharged directly after completing ED service. This distinction is significant because transfer patients are typically more likely to require inpatient admission.

The stochastic primitives are as in Section 2; in particular, arrivals, service times, abandonment behavior, and boarding dynamics are unchanged. We additionally assume that transfer patients are prioritized for service whenever capacity is available, which reflects common operational practice when transfer patients arrive through coordinated pathways.

For implementation, we work with the corresponding fluid approximation. Let $x_i(t)$, $z_i(t)$, and $q_i(t)$ denote the fluid analogues of the number of type-i patients in system, in service, and waiting in queue, respectively, where $q_i(t) = x_i(t) - z_i(t)$, for $i \in \{1,2\}$. Let $b(t)$ denote the boarding fluid. The fluid dynamics are

$$\begin{aligned} \dot{x}_1(t) &= \lambda_1(t) - \mu z_1(t) - \theta q_1(t), \\ \dot{x}_2(t) &= g(t)\lambda_2(t) - \mu z_2(t) - \theta q_2(t), \\ \dot{b}(t) &= (1-p_1)\mu z_1(t) + (1-p_2)\mu z_2(t) - \nu(t) b(t), \end{aligned} \tag{10}$$

where $g(t) \in [0,1]$ denotes the (fluid) transfer acceptance control. Service capacity is allocated according to transfer-priority:

$$\begin{aligned} z_2(t) &= \min\{c(t) - b(t),\, x_2(t)\}, \\ z_1(t) &= \min\{c(t) - b(t) - z_2(t),\, x_1(t)\}. \end{aligned} \tag{11}$$

As in Section 3, the look-ahead decision rule is expressed in terms of a hitting-time functional. Let $H_t(x_1, x_2, b)$ denote the minimum $\Delta > 0$ such that the fluid queues clear by time $t + \Delta$, starting

from (x_1, x_2, b) at time t under full acceptance $g(s) \equiv 1$ for $s \geq t$. We implement a periodic-review policy: at review time $k\delta$, we compute $\hat{H}_{k\delta}$ using the current state $(X_1(k\delta), X_2(k\delta), B(k\delta))$ and the forecasted rates, and apply the same admission decision throughout $[k\delta, (k+1)\delta)$. Concretely, transfer requests are blocked over $[k\delta, (k+1)\delta)$ if

$$\hat{H}_{k\delta}\big(X_1(k\delta), X_2(k\delta), B(k\delta)\big) - \frac{\delta}{2} > -\frac{1}{\theta}\log\left(1 - \frac{l\theta}{h}\right), \tag{12}$$

and accepted otherwise. The shift term $\delta/2$ is a simple correction for within-window state evolution under periodic review.

6.2. Data, Prediction, and Model Calibration

We conduct the case study using detailed ED flow data, including inter-facility transfer requests, from April 2022 to November 2024. The data contain time-stamped records of ED arrival, physician first contact, inpatient bed request, and ED departure, which allow us to reconstruct time-varying arrival patterns, service dynamics, and boarding delays. To support forecasting and calibration, we additionally collect and construct time-varying measures of inpatient occupancy (including non-ED admissions), extract the hospital's recorded transfer-admission status, and merge local weather variables (temperature, precipitation, snowfall, and wind speed). Additional details on data processing are provided in Appendix C.1.

The resulting dataset contains 311,739 ED encounters, of which 20.05% involve a request for inpatient admission. We define boarding time as the duration between inpatient bed request and ED departure. On average, the ED experiences 13.32 arrivals per hour. Boarding times have a mean of 17.88 hours, and their empirical distribution (Figure 7) exhibits substantial dispersion and a pronounced right tail.

Arrival-rate prediction. We forecast arrivals at two-hour resolution using information available at the start of each interval. Predictors include: (i) calendar variables (month, day-of-week, time-of-day, holiday indicators); (ii) lagged arrival counts over the previous one to seven intervals; (iii) a real-time congestion proxy (ED census at the start of the interval); and (iv) contemporaneous weather variables (temperature, precipitation, snowfall, wind, and indicators for extreme heat). To preserve temporal integrity, we split the data chronologically into equally-sized training and test sets and fit both linear regression and XGBoost models. Table 1 reports predictive accuracy; both models achieve similar test-set RMSE (5.69 for linear regression and 5.67 for XGBoost), indicating comparable out-of-sample performance.

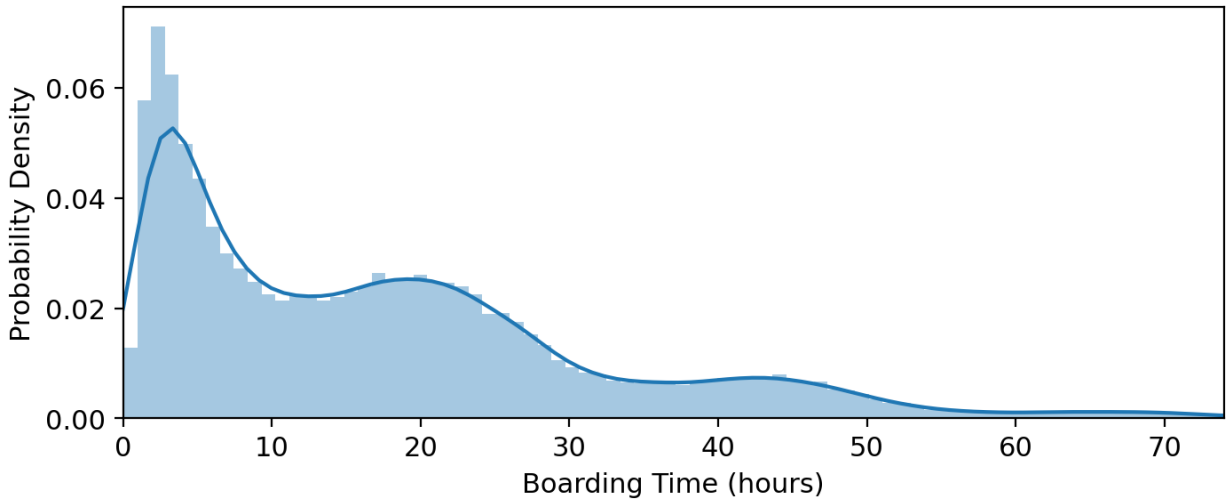

Figure 7 Histogram of boarding time in hours

Dataset	Linear Regression	XGBoost
Train	5.61	4.23
Test	5.69	5.67

Table 1 RMSE values for arrival count prediction models.

Processing-rate estimation and prediction. We next estimate and predict the processing rate governing transfers from the ED to inpatient units. We assume boarding times are exponential, with the processing rate $\nu(t)$ treated as constant within pre-specified 8-hour intervals; specifically, we partition each 24-hour day into three 8-hour periods. For each interval, we observe both completed boarding times and right-censored exposures for patients still boarding at the end of the interval. We estimate $\nu(t)$ by maximum likelihood; the estimator takes the closed form $\widehat{\nu} =$ (#completions)/(total exposure) (see Appendix C.2 for details). The average estimated processing rate is 0.06.

Using these estimates as targets, we train forecasting models for $\nu(t)$ using predictors that include temporal indicators, lagged processing rates, current boarding census, service-level inpatient occupancy, and weather variables. As with arrivals, we split chronologically into equal training and test sets. Table 2 reports predictive accuracy; both linear regression and XGBoost achieve test-set RMSE of approximately 0.02.

Model calibration and simulation. We calibrate a multi-server queueing simulator to reflect observed ED operations over the full study period. Arrivals follow a time-varying Poisson process with piecewise-constant rates over two-hour intervals, estimated directly from the data. The probability that an arriving patient is a transfer patient is estimated as $\gamma = 0.033$. However, based on

Dataset	Linear Regression	XGBoost
Train	0.02	0.01
Test	0.02	0.02

Table 2 RMSE values for processing rate prediction models.

discussions with our clinical partners, transfer requests are likely to be substantially underreported in the administrative data. We therefore also evaluate the policy under higher values of γ. Patient abandonment (patience) times are assumed to follow an exponential distribution, with the abandonment rate estimated via maximum likelihood as $\theta = 0.0347$. Service times are likewise assumed to be exponential, with an estimated mean of 8.1620 hours, corresponding to a service rate $\mu = 1/8.1620$. The number of servers is fixed at $c(t) = 153$, equal to the median number of patients in service during the study period. The probability of direct discharge after ED service completion is estimated separately for external and transfer patients, yielding $p_1 = 0.7928$ and $p_2 = 0.5467$, respectively. Patients who are not discharged directly experience boarding delays governed by the estimated processing rate $\nu(t)$. For numerical stability and to match observed tails, we truncate service times at 84 hours and boarding times at 72 hours, corresponding to the 99.5th percentiles observed in the data.

In implementing the look-ahead policy, we compute the hitting-time functional using fore- casted rates. Specifically,

$$\hat{\lambda}_1(t) = (1-\gamma)\hat{\lambda}(t), \qquad \hat{\lambda}_2(t) = \gamma\hat{\lambda}(t),$$

where $\hat{\lambda}(t)$ is the arrival forecast from the XGBoost model, and $\hat{\nu}(t)$ is the forecasted processing rate from the corresponding XGBoost model. We vary the cost ratio l/h over its feasible range $0 < l/h < 1/\theta$ and consider review intervals δ ranging from one hour to twelve hours. The parameter grid considered is

$$\frac{l}{h} \in \{0, 0.1, 0.5, 1, 4, 8, 12, 16, 20, 24, 26, 28\},\ \delta \in \{1, 2, 4, 6, 8, 12\} \text{ hours},\ \gamma \in \{0.0331, 0.1, 0.2\}. \quad (13)$$

For each configuration, we report the fraction of rejected transfer requests, average waiting time, and average boarding census. To mitigate transient effects, we compute performance after excluding the first and last months of each simulated period.

6.3. Performance Evaluation

The hospital's current policy blocks all transfer requests whenever the boarding census exceeds a fixed threshold $\bar{b}$. During the study period, $\bar{b}$ is approximately 45, and the blocking indicator is active 81.04% of the time. To benchmark this policy, we simulate the system under thresholds $\bar{b} \in \{41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 61, 63, 65\}$. When $\bar{b} = 45$, the simulated blocking fractions are

82.42%, 82.95%, and 78.73% for $\gamma = 3.31\%$, 10%, and 20%, respectively, which are close to the observed blocking frequency of 81.04%.

Figure 8 summarizes performance under the hospital's policy. As $\bar{b}$ increases, average waiting time rises while the fraction of rejected transfer requests declines, reflecting the congestion–access trade-off. The patterns are less pronounced when $\gamma = 3.31\%$, because transfers constitute a small share of arrivals and rejecting them has limited impact on aggregate congestion (additional results are reported in Appendix C.3).

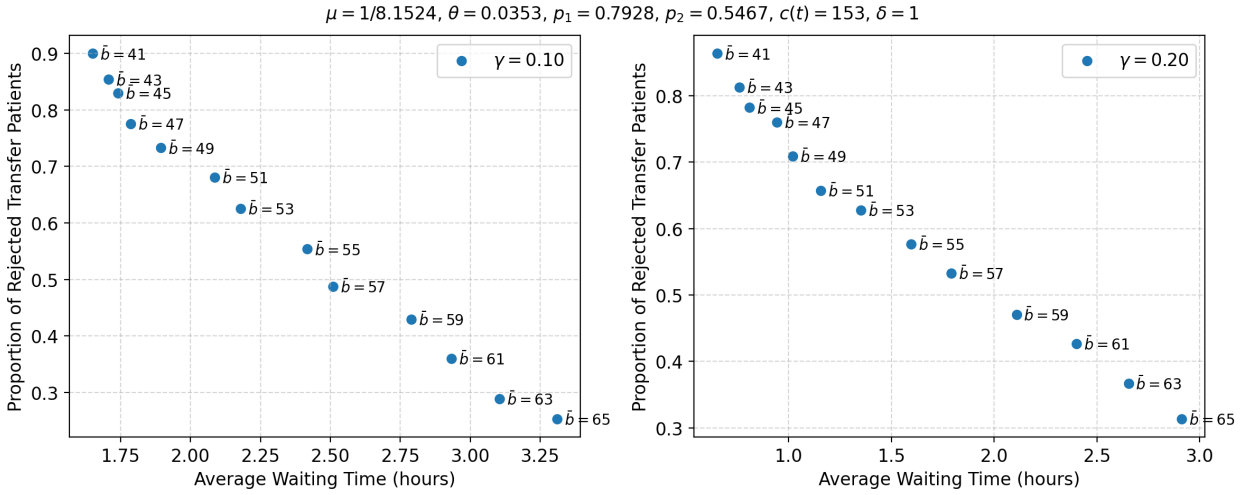

Figure 8 Simulation results under hospital's current policy

We next evaluate the look-ahead policy (12) over the parameter grid in (13) and construct trade-off curves between congestion (average waiting time) and access (transfer rejection rate). Figure 9 reports one such trade-off curve when the review interval $\delta = 1$. As l/h increases, the policy places greater weight on access, rejects fewer transfer requests, and tolerates higher waiting times. Figure 10 reports the analogous trade-off between boarding census and transfer rejection rate. Results for other review intervals are qualitatively similar and are reported in Appendix C.3; consistent with the threshold structure of the policy, performance is relatively insensitive to δ over the range considered.

Figure 11 directly compares the hospital's policy to the look-ahead policy. For both $\gamma = 0.10$ and $\gamma = 0.20$, the look-ahead policy dominates the hospital's policy, achieving lower rejection for comparable (or shorter) waiting times. For example, when $\gamma = 0.20$ and the average waiting time is capped at 2 hours, the look-ahead policy (with $l/h = 12$) reduces the rejection fraction from 0.53

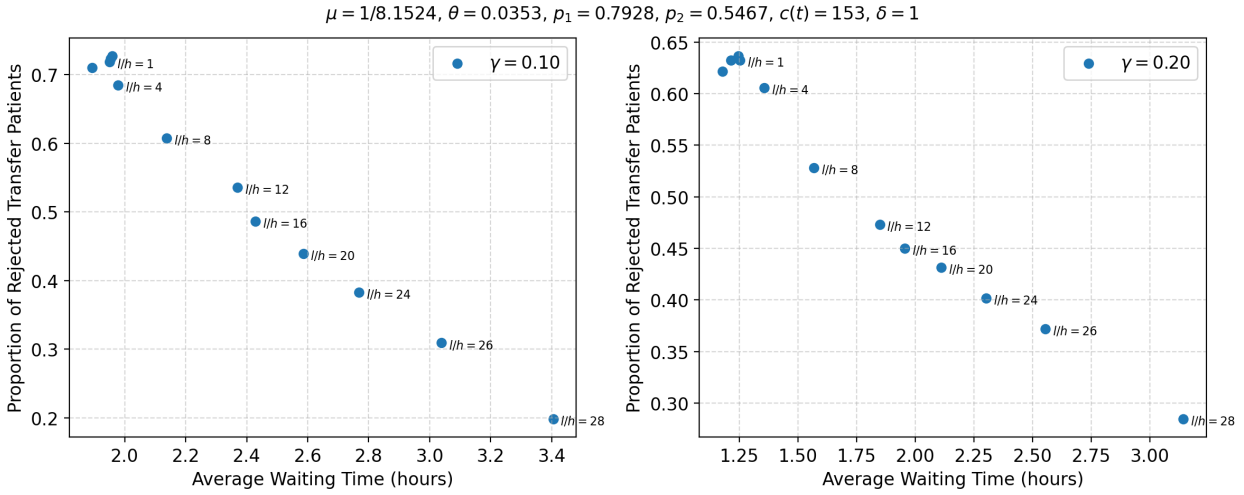

Figure 9 Trade-off between transfer rejection and waiting time under look-ahead policy with $\delta = 1$

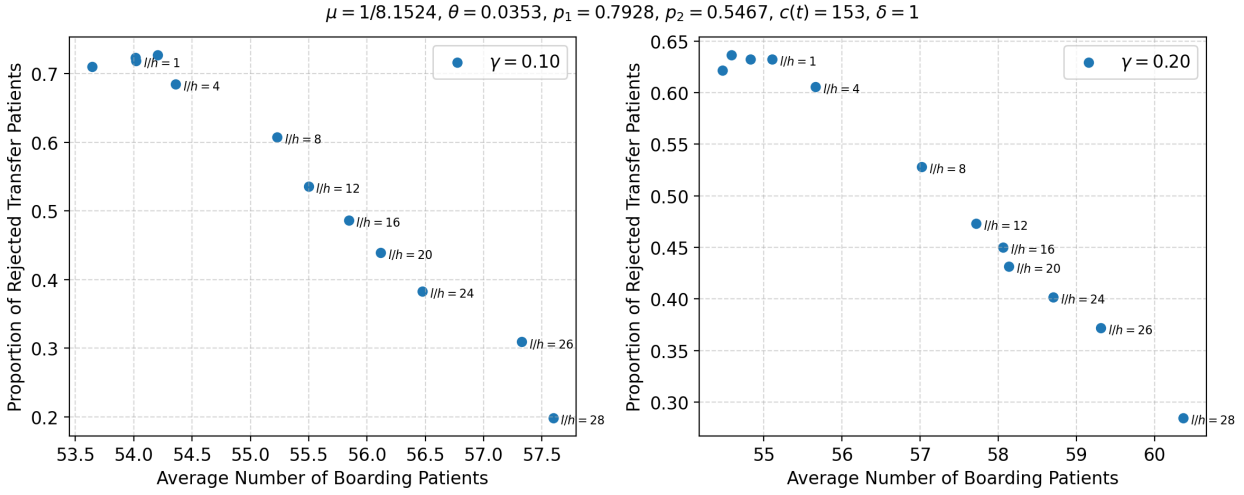

Figure 10 Trade-off between boarding census and waiting time under look-ahead policy with $\delta = 1$

(hospital policy with $\bar{b} = 57$) to 0.47, an 11.22% reduction. Conversely, when the rejection fraction is capped at 0.50, the look-ahead policy reduces average waiting time from 2.11 hours ($\bar{b} = 59$) to 1.85 hours, a 12.33% reduction. These comparisons illustrate that incorporating forecasts into a simple periodic-review rule can substantially improve the congestion–access trade-off relative to current practice.

7. Conclusion

We develop a proactive admission control framework for emergency care settings that integrates predictive information on patient arrivals and inpatient admission delays. As crowding increasingly

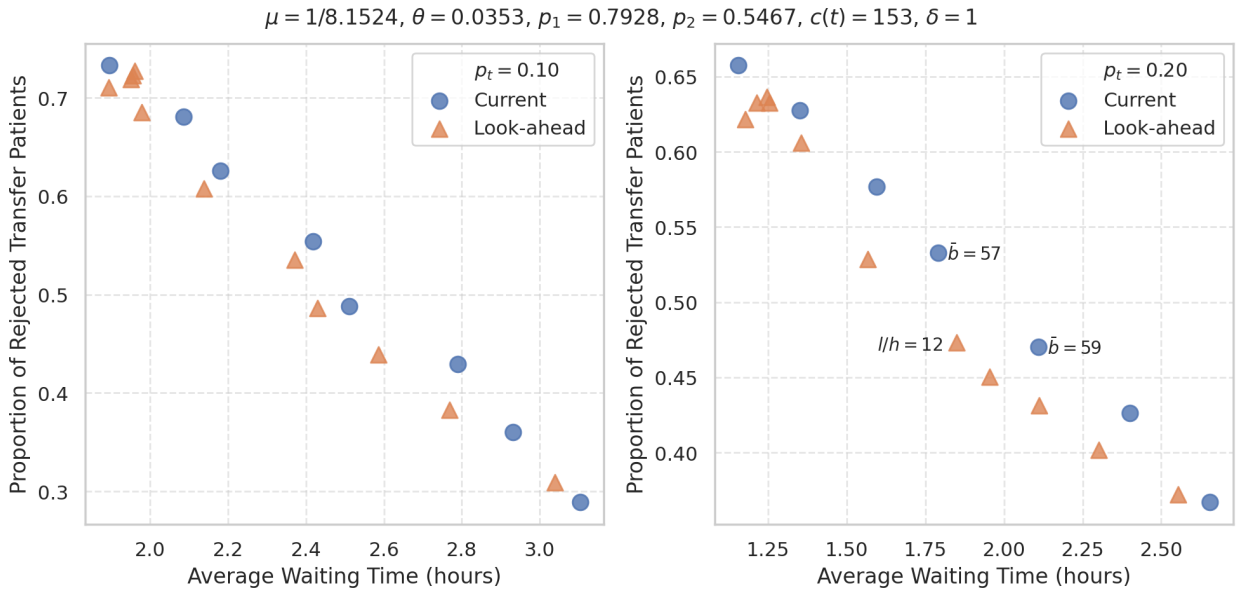

Figure 11 Comparison between Look-ahead Policy versus Current Policy

threatens care quality and contributes to staff burnout, our approach provides a practical, data-driven mechanism for improving patient flow and system performance.

To address the operational complexity of admission decisions under transient congestion, we adopt a fluid approximation and characterize the optimal fluid control. Guided by this analysis, we propose a look-ahead admission policy that is intuitive, easy to implement, and asymptotically optimal as demand and capacity scale.

Extensive synthetic experiments and a data-calibrated case study based on real hospital data demonstrate the effectiveness of the proposed approach. Across a wide range of operating conditions, the look-ahead policy consistently outperforms non-predictive benchmarks by achieving better trade-offs between patient waiting times and transfer rejections.

References

Adepoju, T., Carson, A. L., Jin, H. S., and Manasseh, C. S. (2023). Hospital boarding crises: The impact of urgent vs. prevention responses on length of stay. *Management Science*, 69(10):5948–5963.

Altman, E., Jiménez, T., and Koole, G. (2002). On optimal call admission control in resource-sharing system. *IEEE Transactions on Communications*, 49(9):1659–1668.

Ao, R., Fu, H., and Simchi-Levi, D. (2024). Two-stage online reusable resource allocation: Reservation, overbooking and confirmation call. *arXiv preprint arXiv:2410.15245*.

Asplin, B. R., Magid, D. J., Rhodes, K. V., Solberg, L. I., Lurie, N., and Camargo Jr, C. A. (2003). A conceptual model of emergency department crowding. *Annals of emergency medicine*, 42(2):173–180.

Ata, B. and Peng, X. (2020). An optimal callback policy for general arrival processes: a pathwise analysis. *Operations Research*, 68(2):327–347.

Bassamboo, A., Harrison, J. M., and Zeevi, A. (2005). Dynamic routing and admission control in high-volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems*, 51(3):249–285.

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

Bernstein, S. L., Aronsky, D., Duseja, R., Epstein, S., Handel, D., Hwang, U., McCarthy, M., John McConnell, K., Pines, J. M., Rathlev, N., et al. (2009). The effect of emergency department crowding on clinically oriented outcomes. *Academic Emergency Medicine*, 16(1):1–10.

Bertsimas, D., Pauphilet, J., Stevens, J., and Tandon, M. (2022). Predicting inpatient flow at a major hospital using interpretable analytics. *Manufacturing & Service Operations Management*, 24(6):2809–2824.

Canellas, . s. A. I. G. M. M., Pachamanova, D. A., Perakis, G., Skali Lami, O., and Tsiourvas, A. (2025). A granular approach to optimal and fair patient placement in hospital emergency departments. *Production and Operations Management*, 34(4):575–589.

Centers for Medicare and Medicaid Services (2026). Emergency Medical Treatment and Labor Act (EMTALA).

Delana, K., Savva, N., and Tezcan, T. (2021). Proactive customer service: operational benefits and economic frictions. *Manufacturing & Service Operations Management*, 23(1):70–87.

Ding, Y., Park, E., Nagarajan, M., and Grafstein, E. (2019). Patient prioritization in emergency department triage systems: An empirical study of the canadian triage and acuity scale (ctas). *Manufacturing & Service Operations Management*, 21(4):723–741.

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

Greenwood-Ericksen, M., Kamdar, N., Swenson, K., Pruitt, P., McCrum, M. L., Paul, G., Myaskovsky, L., Kocher, K. E., and Zachrison, K. S. (2025). Emergency department boarding, inpatient census, and interhospital transfer acceptances. *JAMA network open*, 8(5):e2512299–e2512299.

Hartl, R. F., Sethi, S. P., and Vickson, R. G. (1995). A survey of the maximum principles for optimal control problems with state constraints. *SIAM review*, 37(2):181–218.

Heyman, D. P. (1968). Optimal operating policies for m/g/1 queuing systems. *Operations Research*, 16(2):362–382.

Hoot, N. R. and Aronsky, D. (2008). Systematic review of emergency department crowding: causes, effects, and solutions. *Annals of emergency medicine*, 52(2):126–136.

Jacobson, E. U., Argon, N. T., and Ziya, S. (2012). Priority assignment in emergency response. *Operations research*, 60(4):813–832.

Janke, A. T. and Venkatesh, A. K. (2025). Understanding and addressing the US hospital bed shortage-Build, baby, build. *JAMA Network Open*, 8(2):e2460652–e2460652.

Koçağa, Y. L. and Ward, A. R. (2010). Admission control for a multi-server queue with abandonment. *Queueing Systems*, 65(3):275–323.

Lewis, M. E. (2001). Average optimal policies in a controlled queueing system with dual admission control. *Journal of Applied Probability*, 38(2):369–385.

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

Mandelbaum, A., Massey, W. A., and Reiman, M. I. (1998). Strong approximations for markovian service networks. *Queueing Systems*, 30(1):149–201.

McKenna, P., Heslin, S. M., Viccellio, P., Mallon, W. K., Hernandez, C., and Morley, E. J. (2019). Emergency department and hospital crowding: causes, consequences, and cures. *Clinical and experimental emergency medicine*, 6(3):189.

Morley, C., Unwin, M., Peterson, G. M., Stankovich, J., and Kinsman, L. (2018). Emergency department crowding: a systematic review of causes, consequences and solutions. *PloS one*, 13(8):e0203316.

Naor, P. (1969). The regulation of queue size by levying tolls. *Econometrica: journal of the Econometric Society*, pages 15–24.

Ormeci, E. L. (2004). Dynamic admission control in a call center with one shared and two dedicated service facilities. *IEEE Transactions on Automatic Control*, 49(7):1157–1161.

Peng, X. (2024). Admission control to queueing systems with arrival forecast information. *Available at SSRN 4917786*.

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

Seierstad, A. and Sydsæter, K. (1987). *Optimal Control Theory with Economic Applications*. North-Holland, Amsterdam. See Theorem 3.16, Sufficient conditions with nonsmooth functions.

Shi, P., Chou, M. C., Dai, J. G., Ding, D., and Sim, J. (2016). Models and insights for hospital inpatient operations: Time-dependent ed boarding time. *Management Science*, 62(1):1–28.

Silva, D. F., Zhang, B., and Ayhan, H. (2013). Optimal admission control for tandem loss systems with two stations. *Operations Research Letters*, 41(4):351–356.

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

Spencer, J., Sudan, M., and Xu, K. (2014). Queueing with future information. *ACM SIGMETRICS Performance Evaluation Review*, 41(3):40–42.

Stidham Jr, S. (2002). Analysis, design, and control of queueing systems. *Operations Research*, 50(1):197–216.

Sun, B. C., Hsia, R. Y., Weiss, R. E., Zingmond, D., Liang, L.-J., Han, W., McCreath, H., and Asch, S. M. (2013). Effect of emergency department crowding on outcomes of admitted patients. *Annals of emergency medicine*, 61(6):605–611.

Ward, A. R. and Kumar, S. (2008). Asymptotically optimal admission control of a queue with impatient customers. *Mathematics of Operations Research*, 33(1):167–202.

Westfall, J. M. (2024). Transfer as treatment in rural hospitals. *JAMA Network Open*, 7(3):e241845–e241845.

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

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

Zayas-Cabán, G. and Lewis, M. E. (2020). Admission control in a two-class loss system with periodically varying parameters and abandonments. *Queueing Systems*, 94(1):175–210.

Appendix A: Asymptotic Optimality

We focus on the single-surge setting; extensions to multiple surges are analogous but notationally heavier.

We first formalize the planning horizon T in (2). Given the initial state (x_0, b_0), define $T(x_0, b_0)$ as follows. Consider a fluid system that operates at full capacity and accepts all transfer requests, starting from (x_0, b_0). Its dynamics are

$$\dot{x}^o(t) = \lambda_1(t) + \lambda_2(t) - \mu\big(c(t) - b^o(t)\big) - \theta\big(x^o(t) + b^o(t) - c(t)\big), \tag{14}$$

$$\dot{b}^o(t) = (1-p)\mu\big(c(t) - b^o(t)\big) - \nu(t) b^o(t). \tag{15}$$

The planning horizon is the first time (after the surge onset κ) at which congestion clears:

$$T(x_0, b_0) = \inf\big\{t \geq \kappa : \, x^o(t) + b^o(t) \leq c(t) \,\big|\, x^o(0) = x_0, \, b^o(0) = b_0\big\}.$$

We also denote $\bar{V}^*(x_0, b_0)$ as the optimal value of the fluid control problem (5) with initial condition (x_0, b_0).

To establish asymptotic optimality, we consider a sequence of systems indexed by n, in which arrival rates, transfer request rates, and capacity scale linearly with n, while abandonment and service rates remain fixed:

$$\lambda_1^n(t) = n\lambda_1(t), \qquad \lambda_2^n(t) = n\lambda_2(t), \qquad c^n(t) = nc(t).$$

We also scale the initial state as $(X^n(0), B^n(0)) = (nx_0, nb_0)$. Let

$$Z^n(t) = \min\{X^n(t), \, c^n(t) - B^n(t)\}$$

be the number in service and $G^n(t) \in \{0, 1\}$ the transfer acceptance decision. The dynamics satisfy

$$\begin{aligned} X^n(t) &= X^n(0) + A_1\left(\int_0^t n\lambda_1(u)\, du\right) + A_2\left(\int_0^t G^n(u) n\lambda_2(u)\, du\right) - S\left(\int_0^t \mu Z^n(u)\, du\right) - D\left(\int_0^t \theta Q^n(u)\, du\right), \\ B^n(t) &= B^n(0) + S\left(\int_0^t (1-p)\mu Z^n(u)\, du\right) - H\left(\int_0^t \nu(u) B^n(u)\, du\right), \end{aligned} \tag{16}$$

where $Q^n(u) = X^n(u) - Z^n(u)$, and as before, A_1, A_2, S, D, H are independent unit-rate Poisson processes.

Define the fluid-scaled processes

$$\bar{X}^n(t) = \frac{1}{n} X^n(t), \qquad \bar{B}^n(t) = \frac{1}{n} B^n(t), \qquad \bar{Z}^n(t) = \frac{1}{n} Z^n(t).$$

For a policy π^n, the fluid-scaled cost is

$$\bar{V}^{n,\pi^n}(x_0, b_0) = \mathbb{E}\left[\int_0^{T(x_0,b_0)} h(\bar{X}^{n,\pi^n}(t) - \bar{Z}^{n,\pi^n}(t))\, dt + l\, \bar{L}^{n,\pi^n}(T(x_0, b_0))\right], \tag{17}$$

where $\bar{L}^{n,\pi^n} = L^{n,\pi^n}(t)/n$ is the scaled cumulative number of rejected transfers.

For the n-th system, the fluid-scaled objective is

$$\min_{\pi^n} \bar{V}^{n,\pi^n}(x_0, b_0) := \frac{1}{n} V^{n,\pi^n}(nx_0, nb_0)$$

$$=\mathbb{E}\left[\int_0^{T(nx_0,nb_0)} h\big(\bar{X}^{n,\pi^n}(t)-\bar{Z}^{n,\pi^n}(t)\big)dt + l\bar{L}^{n,\pi^n}(T(nx_0,nb_0))\right], \tag{18}$$

where $\bar{L}^{n,\pi^n}(t) = L^{n,\pi^n}(t)/n$.

For state (x,b) at time t, recall that $H_t(x,b)$ is the fluid time-to-clear if all future transfers are accepted (with arrival rate $\lambda_i(t)$, $i=1,2$, processing rate $\nu(t)$, and $c(t)$ servers). Although Theorem 1 characterizes the optimal fluid control through the associated optimality conditions, it is equivalently described by a single switching time: there exists a unique $\zeta^* > \kappa$ such that transfers are rejected for $t < \zeta^*$ and accepted for $t \geq \gamma^*$, where $\gamma^* = \inf\left\{t \geq \kappa : \frac{h}{\theta}\left[1-\exp(-\theta H_t^n(x(t),b(t)))\right] \leq l\right\}$. Motivated by this equivalent switching time characterization, we define the look-ahead policy $\{\bar{\pi}^n\}_{n\geq 1}$ for the stochastic system as follows: transfer requests at time t if $t \geq \zeta^n$ and rejected otherwise, where

$$\zeta^n = \inf\left\{t \geq \kappa : \frac{h}{\theta}\left[1-\exp\big(-\theta H_t(\bar{X}^n(t),\bar{B}^n(t))\big)\right] \leq l\right\}, \tag{19}$$

i.e., the policy switches from rejection to acceptance once the predicted future congestion cost, computed via the fluid time-to-clear from the current state, falls below the blocking cost l.

Our main asymptotic optimality result is stated in the following theorem.

THEOREM 3. *Under Assumption 1, for any sequence of admissible controls* $\{\pi^n\}_{n\geq 1}$,

$$\liminf_{n\to\infty} \bar{V}^{n,\pi^n}(x_0,b_0) \geq \bar{V}^*(x_0,b_0). \tag{20}$$

Moreover, for the sequence of systems operating under the look-ahead policy $\{\bar{\pi}^n\}_{n\geq 1}$,

$$\lim_{n\to\infty} \bar{V}^{n,\bar{\pi}^n}(x_0,b_0) = \bar{V}^*(x_0,b_0). \tag{21}$$

A.1. Proof of Theorem 3

Throughout, all processes are defined on a common probability space supporting the independent unit-rate Poisson processes driving the system.

We will use a functional strong law of large numbers (FSLLN) for Poisson-driven systems with state-dependent Lipschitz rates (e.g., Mandelbaum et al. 1998): if the rates are locally bounded, Lipschitz in the fluid-scaled state, and of linear growth, then the fluid-scaled state process converges almost surely and uniformly on compact time intervals to the unique solution of the corresponding deterministic integral equation. We will apply this result on time intervals where the acceptance rule is constant (all reject or all accept).

Proof of the lower bound (20). Fix any sequence of admissible controls $\{\pi^n\}_{n\geq 1}$.

Taking expectations for (16) and using $\mathbb{E}[N(\Lambda)] = \mathbb{E}[\Lambda]$ for any unit-rate Poisson process N and any nonnegative (possibly random) time change Λ, we obtain

$$\mathbb{E}[X^n(t)] = X^n(0) + \int_0^t n\lambda_1(u)\,du + \int_0^t n\lambda_2(u)\,\mathbb{E}[G^n(u)]\,du - \int_0^t \mu\,\mathbb{E}[Z^n(u)]\,du - \int_0^t \theta\,\mathbb{E}[Q^n(u)]\,du, \tag{22}$$

$$\mathbb{E}[B^n(t)] = B^n(0) + \int_0^t (1-p)\mu\,\mathbb{E}[Z^n(u)]\,du - \int_0^t \nu(u)\,\mathbb{E}[B^n(u)]\,du. \tag{23}$$

Divide by n and define

$$\bar{x}^n(t) := \mathbb{E}[\bar{X}^n(t)], \quad \bar{b}^n(t) := \mathbb{E}[\bar{B}^n(t)], \quad \bar{z}^n(t) := \mathbb{E}[\bar{Z}^n(t)], \quad \bar{g}^n(t) := \mathbb{E}[G^n(t)] \in [0,1].$$

Using $Q^n = X^n - Z^n$, (22)–(23) become

$$\bar{x}^n(t) = x_0 + \int_0^t \Big(\lambda_1(u) + \lambda_2(u)\bar{g}^n(u) - \mu\bar{z}^n(u) - \theta(\bar{x}^n(u) - \bar{z}^n(u))\Big)du, \tag{24}$$

$$\bar{b}^n(t) = b_0 + \int_0^t \Big((1-p)\mu\bar{z}^n(u) - \nu(u)\bar{b}^n(u)\Big)du. \tag{25}$$

Moreover, feasibility implies, for each t,

$$0 \le \bar{Z}^n(t) = \frac{1}{n}\min\{X^n(t),\, nc(t) - B^n(t)\} \le \min\{\bar{X}^n(t),\, c(t) - \bar{B}^n(t)\}.$$

Taking expectations and using Jensen's inequality for the concave map $\min\{\cdot,\cdot\}$ gives

$$0 \le \bar{z}^n(t) \le \mathbb{E}\big[\min\{\bar{X}^n(t),\, c(t) - \bar{B}^n(t)\}\big] \le \min\{\bar{x}^n(t),\, c(t) - \bar{b}^n(t)\}. \tag{26}$$

Therefore, $(\bar{x}^n, \bar{b}^n)$ together with controls $(\bar{z}^n, \bar{g}^n)$ is an admissible solution of the (relaxed) fluid dynamics.

Next, consider the n-th fluid-scaled cost on $[0,T]$:

$$\bar{V}^{n,\pi^n}(x_0,b_0) = \mathbb{E}\left[\int_0^T h(\bar{X}^n(t) - \bar{Z}^n(t))\,dt + \int_0^T l\,\lambda_2(t)\,(1 - G^n(t))\,dt\right],$$

where we used the fact that the scaled cumulative rejection count satisfies $\mathbb{E}[\bar{L}^n(T)] = \int_0^T \lambda_2(t)\,\mathbb{E}[1 - G^n(t)]dt$.

Using Jensen's inequality and linearity of expectation,

$$\begin{aligned}\bar{V}^{n,\pi^n}(x_0,b_0) &\ge \int_0^T h\big(\mathbb{E}[\bar{X}^n(t) - \bar{Z}^n(t)]\big)\,dt + \int_0^T l\,\lambda_2(t)\,(1 - \mathbb{E}[G^n(t)])\,dt \\ &= \int_0^T \Big[h(\bar{x}^n(t) - \bar{z}^n(t)) + l\,\lambda_2(t)(1 - \bar{r}^n(t))\Big]dt.\end{aligned}$$

Since $(\bar{x}^n, \bar{b}^n, \bar{z}^n, \bar{g}^n)$ is feasible fluid solution and $\bar{V}^*(x_0,b_0)$ is the infimum over all feasible controls, we have

$$\int_0^T \Big[h(\bar{x}^n(t) - \bar{z}^n(t)) + l\,\lambda_2(t)(1 - \bar{r}^n(t))\Big]dt \;\ge\; \bar{V}^*(x_0,b_0).$$

Thus, for every n, $\bar{V}^{n,\pi^n}(x_0,b_0) \ge \bar{V}^*(x_0,b_0)$, and taking lim inf yields (20).

Proof of (21). Now consider the look-ahead policy $\{\bar{\pi}^n\}$ defined by (19). Let $(x^*(t), b^*(t), z^*(t), r^*(t))$ be the optimal fluid solution, where $r^*(t) = \mathbf{1}\{t \ge \zeta^*\}$ with ζ^* the unique fluid switching time.

We prove that under $\bar{\pi}^n$,

$$\sup_{0\le t\le T}\big|(\bar{X}^n(t), \bar{B}^n(t)) - (x^*(t), b^*(t))\big| \xrightarrow[n\to\infty]{a.s.} 0, \tag{27}$$

and then show convergence of costs.

Step 1: FSLLN on intervals where the acceptance decision is constant. For any deterministic $s \in [0,T]$, on an interval $[0,s]$ where $G^n(t)$ is fixed equal to 0, the arrival rates and departure rates of (X^n, B^n) can be written as Lipschitz functions of the fluid-scaled state $(\bar{X}^n, \bar{B}^n)$ with linear growth, because

$$\bar{Z}^n(t) = \min\{\bar{X}^n(t),\, c(t) - \bar{B}^n(t)\}$$

is Lipschitz in $(\bar{X}^n, \bar{B}^n)$ and all primitive rates are locally bounded. Hence the FSLLN applies on such intervals, yielding almost sure u.o.c. convergence of $(\bar{X}^n, \bar{B}^n)$ to the corresponding deterministic fluid trajectory driven by the same fixed reject decision.

Step 2: Convergence of the switching times $\zeta^n \to \zeta^*$. Define the fluid look-ahead statistic

$$\Psi(t;x,b) := \frac{h}{\theta}\Big(1 - e^{-\theta H_t(x,b)}\Big),$$

and its stochastic analog

$$\Psi^n(t) := \frac{h}{\theta}\Big(1 - e^{-\theta H_t(\bar{X}^n(t), \bar{B}^n(t))}\Big).$$

Recall that the switching time is defined as first passage time:

$$\zeta^* = \inf\{t \geq \kappa : \Psi(t) \leq l\}, \quad \zeta^n = \inf\{t \geq \kappa : \Psi^n(t) \leq l\}.$$

Fix $\varepsilon > 0$. We rely on the transversality of the fluid crossing. Under Assumption 1, $\Psi(t)$ is continuous and crosses l strictly at ζ^*. Thus, there exists $\delta > 0$ such that:

$$\inf_{t \in [\kappa, \zeta^* - \varepsilon]} \Psi(t) \geq l + \delta. \tag{28}$$

Consider the interval $[0, \zeta^* - \varepsilon]$. On this interval, the optimal fluid control is constant: $r^*(t) = 0$ (Reject). Let $(\tilde{x}(t), \tilde{b}(t))$ be the deterministic fluid trajectory strictly following the Reject policy. By uniqueness of the fluid ODE solutions, on this specific interval, $(x^*(t), b^*(t)) \equiv (\tilde{x}(t), \tilde{b}(t))$. By Step 1 (FSLLN with fixed policy), the stochastic system driven by the Reject policy, denoted by $(\tilde{X}^n, \tilde{B}^n)$, converges u.o.c. to $(\tilde{x}, \tilde{b})$. By the continuity of the mapping from state to statistic,

$$\sup_{0 \leq t \leq \zeta^* - \varepsilon} |\tilde{\Psi}^n(t) - \Psi(t)| \xrightarrow[n\to\infty]{a.s.} 0,$$

where $\tilde{\Psi}^n$ is the statistic associated with the "Reject-only" trajectory. Combining this with (28), for sufficiently large n, almost surely:

$$\tilde{\Psi}^n(t) > l \quad \text{for all } t \in [\kappa, \zeta^* - \varepsilon].$$

This implies that the look-ahead policy, which chooses Reject when $\Psi^n(t) > l$, will indeed choose Reject throughout $[\kappa, \zeta^* - \varepsilon]$. Consequently, the actual process matches the "Reject-only" process $(\bar{X}^n(t) = \tilde{X}^n(t))$ on this interval, and no switching occurs. Thus, $\liminf_{n\to\infty} \zeta^n \geq \zeta^* - \varepsilon$ a.s.

For the other direction, we argue by contradiction. Suppose there exists $\varepsilon > 0$ such that the event $E^n := \{\zeta^n > \zeta^* + \varepsilon\}$ occurs infinitely often. On the event E^n, the stochastic policy maintains the Reject action throughout the interval $[0, \zeta^* + \varepsilon]$. Recall that $(\tilde{x}(t), \tilde{b}(t))$ denotes the deterministic fluid trajectory driven by the policy that always rejects, i.e., $g(t) \equiv 0$. Since the optimal fluid policy also rejects on $[0, \zeta^*)$, uniqueness of ODE solutions implies $(\tilde{x}(t), \tilde{b}(t)) = (x^*(t), b^*(t))$ for all $t \leq \zeta^*$. For $t > \zeta^*$, the optimal policy accepts, whereas the always-reject policy continues to reject.

Continuing to reject yields a state at time $\zeta^*+\varepsilon$ no larger than switching to accept from ζ^* onward; consequently,

$$\Psi(\zeta^*+\varepsilon;\tilde{x}(\zeta^*+\varepsilon),\tilde{b}(\zeta^*+\varepsilon)) \le \Psi(\zeta^*+\varepsilon;x^*(\zeta^*+\varepsilon),b^*(\zeta^*+\varepsilon)).$$

Moreover, by Assumption 1, the optimal trajectory satisfies $\Psi(\zeta^*+\varepsilon;x^*(\zeta^*+\varepsilon),b^*(\zeta^*+\varepsilon)) \le l-\delta'$ for some $\delta'>0$. Hence,

$$\Psi(\zeta^*+\varepsilon;\tilde{x}(\zeta^*+\varepsilon),\tilde{b}(\zeta^*+\varepsilon)) \le l-\delta'.$$

By the FSLLN (Step 1), on the event E^n, the stochastic state converges to the "always Reject" fluid trajectory on $[0,\zeta^*+\varepsilon]$. Continuity of Ψ then implies

$$\Psi^n(\zeta^*+\varepsilon) \xrightarrow{a.s.} \Psi(\zeta^*+\varepsilon;\tilde{x}(\zeta^*+\varepsilon),\tilde{b}(\zeta^*+\varepsilon)) \le l-\delta'.$$

Thus, for sufficiently large n, we have $\Psi^n(\zeta^*+\varepsilon)<l$. By definition, ζ^n is the first time the statistic drops below l. The fact that the condition is already satisfied at $\zeta^*+\varepsilon$ contradicts the assumption that $\zeta^n>\zeta^*+\varepsilon$ on E^n. Thus, $\limsup_{n\to\infty}\zeta^n \le \zeta^*+\varepsilon$ a.s.

Since ε is arbitrary, we conclude $\zeta^n\to\zeta^*$ almost surely.

Step 3: Stitching the fluid limits across the (asymptotically matching) switching times. Define $\tau_-^n := \zeta^n\wedge\zeta^*$ and $\tau_+^n := \zeta^n\vee\zeta^*$. On $[0,\tau_-^n]$, both the stochastic policy and the fluid optimal policy reject transfers. The FSLLN in Step 1 implies

$$\sup_{0\le t\le\tau_-^n}\left|(\bar{X}^n(t),\bar{B}^n(t))-(x^*(t),b^*(t))\right| \xrightarrow[n\to\infty]{a.s.} 0.$$

Next, we handle the transition. By Step 2, $\tau_+^n-\tau_-^n\to 0$ a.s. Using the local boundedness of the transition rates, the state cannot change macroscopically during this vanishing interval. Thus,

$$(\bar{X}^n(\tau_+^n),\bar{B}^n(\tau_+^n)) \xrightarrow[n\to\infty]{a.s.} (x^*(\zeta^*),b^*(\zeta^*)).$$

On the remaining interval $[\tau_+^n,T]$, both policies accept transfers ($r=1$). Let $(\hat{x}^n(t),\hat{b}^n(t))$ denote the unique solution to the Accept fluid ODE starting at time τ_+^n with initial state $(\bar{X}^n(\tau_+^n),\bar{B}^n(\tau_+^n))$. We decompose the error using the triangle inequality:

$$|(\bar{X}^n(t),\bar{B}^n(t))-(x^*(t),b^*(t))| \le \underbrace{|(\bar{X}^n(t),\bar{B}^n(t))-(\hat{x}^n(t),\hat{b}^n(t))|}_{\text{(A)}} + \underbrace{|(\hat{x}^n(t),\hat{b}^n(t))-(x^*(t),b^*(t))|}_{\text{(B)}}.$$

Term (A) vanishes almost surely by applying the FSLLN (as in Step 1) to the process starting at τ_+^n. Term (B) vanishes by the continuous dependence of ODE solutions on initial state Combining these yields

$$\sup_{\tau_+^n\le t\le T}\left|(\bar{X}^n(t),\bar{B}^n(t))-(x^*(t),b^*(t))\right| \xrightarrow[n\to\infty]{a.s.} 0.$$

Combining the convergence on $[0,\tau_-^n]$ and $[\tau_+^n,T]$ proves (27).

Step 4: Convergence of costs. First, the function $\bar{Z}^n(t)=\min\{\bar{X}^n(t),c(t)-\bar{B}^n(t)\}$ is continuous in the state $(\bar{X}^n,\bar{B}^n)$. Given the uniform convergence established in (27), we have

$$\sup_{0\le t\le T}|\bar{Z}^n(t)-z^*(t)| \xrightarrow[n\to\infty]{a.s.} 0.$$

Next, the rejection indicator under $\bar{\pi}^n$ is $1-G^n(t)=\mathbf{1}\{t<\zeta^n\}$, so

$$\int_0^T \lambda_2(t)(1-G^n(t))dt=\int_0^{\zeta^n}\lambda_2(t)\,dt \xrightarrow[n\to\infty]{a.s.} \int_0^{\zeta^*}\lambda_2(t)\,dt,$$

by $\zeta^n\to\zeta^*$ and boundedness of $\lambda_2(\cdot)$.

Finally, since the fluid-scaled states are uniformly integrable (due to being dominated by the square-integrable arrival process over the finite horizon $[0,T]$), we can apply the dominated convergence theorem to exchange the limit and the expectation. Combining this with the almost sure convergence of the pathwise costs derived above, we obtain

$$\lim_{n\to\infty}\bar{V}^{n,\bar{\pi}^n}(x_0,b_0)=\int_0^T h\big(x^*(t)-z^*(t)\big)dt+\int_0^T l\,\lambda_2(t)\big(1-r^*(t)\big)dt=\bar{V}^*(x_0,b_0),$$

establishing (21). Q.E.D.

Appendix B: Proof of the Fluid Optimal Control (Theorems 1–2)

In this section, we prove the optimal fluid admission control policy via Pontryagin's Minimum Principle. We first state a version of Pontryagin's Minimum Principle adapted to (5). We then verify the (sufficient) conditions of Pontryagin's Minimum Principle.

Fix any admissible $g(\cdot)$. It is without loss of optimality to enforce work conservation, which implies the service rate $z(t)$ must satisfy:

$$z(t)=\min\{x(t),c(t)-b(t)\}. \tag{29}$$

Equivalently, this can be formulated as treating $z(t)$ as an auxiliary variable subject to the constraints:

$$z(t)\le x(t),\quad z(t)\le c(t)-b(t),\quad \text{and } z(t)=c(t)-b(t) \text{ if } x(t)>c(t)-b(t).$$

For the purpose of deriving the optimal control, we define the **Hamiltonian** $\mathcal{H}$ representing the system dynamics and costs, and the **Augmented Lagrangian** $\mathcal{L}$ which incorporates the boundary constraints via multipliers.

Define the Hamiltonian $\mathcal{H}:[0,\tau]\times\mathbb{R}^2\times\mathbb{R}\times\mathbb{R}^2\to\mathbb{R}$ as:

$$\begin{aligned}\mathcal{H}(t,x,b,g,p_x,p_b):=\quad & p_x\Big[\lambda_1(t)+g\lambda_2(t)-\mu z-\theta(x-z)\Big]+p_b\Big[(1-p)\mu z-\nu(t)b\Big]\\ & +h(x-z)+l\lambda_2(t)(1-g).\end{aligned} \tag{30}$$

To handle the nonsmooth dynamics $z=\min\{x,c-b\}$, we introduce the Lagrangian multipliers $\eta_1(t),\eta_2(t)\ge 0$ corresponding to the capacity and queue constraints, respectively. The Augmented Lagrangian is:

$$\mathcal{L}(t,x,b,g,z,p,\eta):=\mathcal{H}-\eta_1\big(c(t)-b-z\big)-\eta_2(x-z). \tag{31}$$

The following sufficiency theorem is adapted from standard results on optimal control with state constraints (e.g., (Seierstad and Sydsæter 1987, Hartl et al. 1995)).

THEOREM 4. *Consider the optimal control problem (5). Let (x^*,b^*,g^*,z^*) be a feasible trajectory. Suppose there exist absolutely continuous adjoint functions $p_x(\cdot),p_b(\cdot)$ and piecewise continuous multipliers $\eta_1(\cdot),\eta_2(\cdot)\ge 0$ defined on $[0,\tau]$ such that for almost all t:*

1. ***Convexity:*** *The map* $(x,b,g)\mapsto \mathcal{H}\big(t,x,b,g,p_x,p_b\big)$ *is convex on* $\mathbb{R}^2\times[0,1]$.
2. ***Adjoint Equations:*** *The time derivatives of the adjoint variables satisfy:*
$$-\dot{p}_x(t)=\frac{\partial\mathcal{L}}{\partial x},\qquad -\dot{p}_b(t)=\frac{\partial\mathcal{L}}{\partial b},$$
evaluated at the optimal solution.
3. ***Stationarity & Complementarity:*** *The auxiliary variable* z^* *maximizes the Lagrangian, and the multipliers satisfy complementary slackness:*
$$\frac{\partial\mathcal{L}}{\partial z}=0,\quad \eta_1(c-b^*-z^*)=0,\quad \eta_2(x^*-z^*)=0.$$
4. ***Minimization Condition:***
$$g^*(t)\in\arg\min_{0\le g\le 1}\mathcal{H}\big(t,x^*(t),b^*(t),g,p_x(t),p_b(t)\big).$$
5. ***Transversality:***
$$p_x(\tau)=0,\quad p_b(\tau)=0.$$

Then $(x^*(\cdot),b^*(\cdot),g^*(\cdot))$ *is an optimal solution.*

B.1. Proof of Theorem 1

We prove the theorem by verifying the conditions in Theorem 4.

Lagrangian Multipliers Construction. We define the multipliers piecewise: For $t<\tau$, the capacity constraint is active ($z=c-b$). We set:

$$\eta_2(t)=0,\qquad \eta_1(t)=h+(\mu-\theta)p_x(t)-(1-p)\mu p_b(t).$$

For $t\ge\tau$, the queue is empty ($z=x$). We set:

$$\eta_1(t)=0,\qquad \eta_2(t)=-(\mu-\theta)p_x(t)+(1-p)\mu p_b(t)\quad(\text{Note: } p_x=0\implies \eta_2\approx(1-p)\mu p_b).$$

Adjoint Construction. We construct the adjoint trajectories by integrating backward from τ, using the transversality conditions $p_x(\tau)=0, p_b(\tau)=0$.

p_x**:** For $t<\tau$, the adjoint equation is $-\dot{p}_x=\frac{\partial\mathcal{L}}{\partial x}$. Using the multipliers defined above (where $\eta_2=0$), this reduces to $-\dot{p}_x=\theta p_x-h$. The solution is:

$$p_x(t)=\frac{h}{\theta}\Big(1-e^{-\theta(\tau-t)}\Big).$$

For $t\ge\tau$, we have $p_x(t)=0$, which satisfies the empty-state adjoint inclusion with the constructed η_4.

p_b**:** For $t<\tau$, the adjoint equation is $-\dot{p}_b=\frac{\partial\mathcal{L}}{\partial b}$. With η_1 active, this becomes $-\dot{p}_b=\eta_1-\nu p_b$. Substituting η_1, we get the linear ODE:

$$-\dot{p}_b(t)=\Big[h+(\mu-\theta)p_x(t)-(1-p)\mu p_b(t)\Big]-\nu p_b(t).$$

Solving backward from $p_b(\tau)=0$ yields a strictly positive trajectory for $p_b(t)$.

These explicit constructions satisfy the Adjoint Inclusion and Transversality conditions globally. Furthermore, since $p_b(t)$ accumulates the positive cost h, we have $\eta_1(t)>0$, ensuring the Hamiltonian is globally convex. Furthermore, the specific construction of the multipliers ensures the stationarity condition $\frac{\partial\mathcal{L}}{\partial z}=0$ is satisfied.

Minimization Condition. Let K be the threshold value associated with the switching condition in (8):

$$\frac{h}{\theta}\left(1-e^{-\theta K}\right)=l \quad \Longrightarrow \quad K=-\frac{1}{\theta}\ln\left(1-\frac{l\theta}{h}\right).$$

The proposed policy can be rewritten as:

$$g^*(t)=\begin{cases}1 & \text{if } H_t(x(t),b(t))\le K,\\ 0 & \text{if } H_t(x(t),b(t))>K.\end{cases}$$

By definition, if $g(t)\equiv 1$, then the time derivative of the hitting time is $\frac{d}{dt}H_t(x(t),b(t))=-1$. If we reject ($g=0$), the queue receives fewer arrivals, so it drains faster than the hypothetical rate. Thus, along a trajectory with $g=0$, we have $\frac{d}{dt}H_t(x(t),b(t))<-1$. This implies that if the system starts with $H_t>K$ (Reject region), H_t will strictly decrease until it hits K. Once it enters the region $H_t\le K$ (Accept region), the control switches to $g=1$, and H_t decreases at rate -1 until it reaches 0. Therefore, the optimal trajectory has a single switch at time ζ: it starts with $g=0$ for the interval $[0,\zeta)$, and then switches to $g=1$ for $[\zeta,\tau]$.

Phase A: The Acceptance Phase ($t\ge\zeta$). In this phase, the optimal policy is $g=1$. Since all transfer requests are accepted, the actual dynamics match the "hypothetical" dynamics used to define H_t. Therefore, the remaining time to clear is exactly the hitting time, i.e., $\tau-t=H_t(x(t),b(t))$. Substituting this into the adjoint expression:

$$p_x(t)=\frac{h}{\theta}\left(1-e^{-\theta(\tau-t)}\right)=\frac{h}{\theta}\left(1-e^{-\theta H_t(x,b)}\right).$$

Since $t\ge\zeta$ implies $H_t\le K$, we have

$$p_x(t)\le\frac{h}{\theta}(1-e^{-\theta K})=l.$$

The minimization condition implies that when $p_x(t)\le l$, choosing $g^*=1$ minimizes the Hamiltonian. Thus, the policy is optimal in this phase.

Phase B: The Rejection Phase ($t<\zeta$). By construction of the switching time ζ, we have $\tau-\zeta=K$. For any time $t<\zeta$, $\tau-t>\tau-\zeta=K$. Then,

$$p_x(t)=\frac{h}{\theta}\left(1-e^{-\theta(\tau-t)}\right)>\frac{h}{\theta}\left(1-e^{-\theta K}\right)=l.$$

The minimization condition implies that when $p_x(t)>l$, choosing $g^*=0$ is optimal. Thus, the policy is optimal in this phase. Q.E.D.

B.2. Proof of Theorem 2

We prove the case where $q^*(\tau_b)=0$ only. The proof for the case where the queue never empties between surges, i.e., $q^*(\kappa_b)=0$ follows from the proof of Theorem 1.

We start by defining a few things:

- **Congested Phase:** Defined as the set of times $\mathcal{T}_{busy}=\{t\mid x^*(t)+b^*(t)>c(t)\}$. In this region, $q^*(t)>0$, so the active constraint is $z=c-b$.
- **Uncongested Phase:** Defined as $\mathcal{T}_{free}=\{t\mid x^*(t)+b^*(t)\le c(t)\}$. In this region, $q^*(t)=0$, so the active constraint is $z=x$.

We define three distinct intervals based on the optimal trajectory:

1. **First Busy Phase ($[0, \tau_1)$):** The initial backlog clears at τ_1.
2. **Uncongested Phase ($[\tau_1, \kappa_{cong})$):** This interval includes the "Gap" (low demand) and the "Ramp-Up" (high demand filling capacity) if λ jumps before the queue becomes positive.
3. **Second Busy Phase ($[\kappa_{cong}, \tau_2)$):** Congestion reforms at κ_{cong} and clears finally at τ_2.

We next construct the adjoint variables $p_x(t)$ and $p_b(t)$ by integrating backward from τ_2. Let $S(t) = h + (\mu - \theta)p_x(t)$ be the "congestion source term" and let the integrating factor be $E(t,s) = \exp\left(-\int_t^s [(1-p)\mu + \nu(u)]du\right)$.

$$p_b(t) = \begin{cases} \int_t^{\tau_2} E(t,s)S(s)ds, & t \in [\kappa_{cong}, \tau_2) \\ p_b(\kappa_{cong}) \exp\left(-\int_t^{\kappa_{cong}} \nu(u)du\right), & t \in [\tau_1, \kappa_{cong}) \\ \int_t^{\tau_1} E(t,s)S(s)ds + p_b(\tau_1)E(t,\tau_1), & t \in [0, \tau_1) \\ 0, & t \geq \tau_2 \end{cases}$$

$$p_x(t) = \begin{cases} \frac{h}{\theta}(1 - e^{-\theta(\tau_2 - t)}), & t \in [\kappa_{cong}, \tau_2) \\ e^{-\mu(\kappa_{cong} - t)}p_x(\kappa_{cong}) + \int_t^{\kappa_{cong}} e^{-\mu(s-t)}(1-p)\mu p_b(s)\, ds, & t \in [\tau_1, \kappa_{cong}) \\ \frac{h}{\theta}(1 - e^{-\theta(\tau_1 - t)}) + p_x(\tau_1)e^{-\theta(\tau_1 - t)}, & t \in [0, \tau_1) \\ 0, & t \geq \tau_2 \end{cases}$$

The state constraint multipliers are defined as:

$$\eta_1(t) = \begin{cases} S(t) - (1-p)\mu p_b(t), & t \in \mathcal{T}_{busy} \\ 0, & t \in \mathcal{T}_{free} \end{cases}$$

$$\eta_2(t) = \begin{cases} 0, & t \in \mathcal{T}_{busy} \\ S(t) - (1-p)\mu p_b(t), & t \in \mathcal{T}_{free} \end{cases}$$

We next verify the conditions for the constructed solution.

Adjoint Equations ($-\dot{p} = \nabla L$): In Busy Periods: ($\eta_2 = 0$): Differentiating $p_1^*(t)$ yields $-\dot{p}_1 = \theta p_1 - h$, which matches $\frac{\partial L}{\partial x}$. Differentiating $p_2^*(t)$ yields $-\dot{p}_2 = S(t) - [(1-p)\mu + \nu]p_2$. Substituting $S(t)$ and η_3, this matches $\frac{\partial L}{\partial b} = \eta_3 - \nu p_2$.

In Uncongested Phase ($\eta_1 = 0$): Differentiating $p_1^*(t)$ yields $-\dot{p}_1 = -\mu p_1 + (1-p)\mu p_2$. The Lagrangian derivative is $\frac{\partial L}{\partial x} = (h - \theta p_1) - \eta_4$. Substituting $\eta_4 = h + (\mu - \theta)p_1 - (1-p)\mu p_2$, we confirm the equation holds.

Stationarity ($\frac{\partial L}{\partial z} = 0$):

$$\frac{\partial L}{\partial z} = \underbrace{-h + (\theta - \mu)p_x + (1-p)\mu p_b}_{-\frac{\partial H}{\partial z}} + \eta_1 + \eta_2.$$

Using our definitions, $\eta_1 + \eta_2 = h + (\mu - \theta)p_x - (1-p)\mu p_b$, so the terms sum to zero in both phases. Positivity $\eta_1(t) > 0$ in the busy period follows from $-\dot{p}_2 > 0$ (accumulation of positive cost backward in time), ensuring global convexity of the Hamiltonian.

Minimization: The Hamiltonian is minimized when $g^*(t)=1 \iff p_1^*(t)\le l$. The Pontryagin minimization condition requires selecting $g^*(t)$ to minimize the linear term in the Hamiltonian:

$$g^*(t)=\begin{cases}1 & \text{if } p_1^*(t)\le l,\\ 0 & \text{if } p_1^*(t)>l.\end{cases}$$

We define the switching function $\sigma(t):=p_1^*(t)-l$. We now prove that the sign of $\sigma(t)$ coincides exactly with the sign of $H_t(x^*(t),b^*(t))-K$.

In Busy Periods:

$$p_x(t)=\frac{h}{\theta}\left(1-e^{-\theta(\tau_{clear}-t)}\right),$$

where τ_{clear} is the clearing time of the current surge. The proof follows the same lines of argument as Theorem 1

In Uncongested Period: By construction, $p_x(t)\le p_x(\kappa_{\text{cong}})\le l$ for all $t\in[\tau_1,\kappa_{\text{cong}})$. Q.E.D.

Appendix C: Case Study

C.1. Data Processing

For transfer patients data, we exclude all ED-to-inpatient and inpatient-to-inpatient transfer requests, as these do not constitute ED arrivals. All ED-to-ED transfer requests are included regardless of whether the requests are accepted or declined; these account for 36.04% of all documented transfer requests. Together with non-transfer ED patients (eternal arrivals), they constitute our study cohort. We further exclude patients requesting inpatient services outside of general medicine, oncology, cardiology, surgery, neurology, or orthopedics, which accounts for 3.20% of ED patients. We include all admitted inpatients when computing hourly inpatient occupancy for each of the medical services listed above. Service-level inpatient occupancy is controlled for when predicting processing rates. In addition, we control for local hourly weather conditions (temperature, precipitation, snowfall, and wind speed) in all predictive analyses. Finally, to ensure the accuracy of occupancy and census measures, we exclude observations from the initial and terminal months of the data. The resulting final cohort consists of 311,739 ED encounters. This case study was conducted with relevant Institutional Review Board approval.

C.2. Estimate Processing Rate

To estimate the processing rate of patient boarding from the ED to the inpatient side, we assume the boarding times B following an exponential distribution, where the processing rate ν_k is assumed constant within each pre-specified time interval k. For each 8-hour interval $[t_k,t_{k+1})$, the data consisted of every patient i who have some presence of boarding in focal interval k, including two types of observations: (1) boarding times for patients who completed the boarding process within the focal interval, and (2) truncated boarding times for patients who remained in boarding status at the end of the interval. The likelihood function for ν_k is constructed by combining the probabilities of these two observation types. Let b_i and d_i be the boarding start and end time for patient i respectively. For a completed boarding observation with time $B_i=d_i-\max(b_i,t_k)$, the probability density function of the exponential distribution is:

$$f(B_i)=\nu_k e^{-\nu_k B_i}.$$

For a truncated observation, where the boarding time exceeds $B_i=t_{k+1}-\max(b_i,t_k)$, the survival function is given by:

$$\mathbb{P}(B>B_i)=e^{-\nu_k B_i}.$$

Assuming there are n patients in the interval, with n_c completed observations and n_t truncated observations ($n_c + n_t = n$), the likelihood function is expressed as:

$$L(\nu_k) = \prod_{i=1}^{n_c} \nu_k e^{-\nu B_i} \prod_{j=1}^{n_t} e^{-\nu_k B_j}.$$

We used the Maximum Likelihood Estimation approach to estimate $\nu(t)$ for each interval by maximizing the log-likelihood function. In particular,

$$\nu_k = \frac{n_c}{\sum_{i=1}^{n} B_i}, \quad \nu(t) = \nu_k \text{ for } t \in [t_k, t_{k+1})$$

This allowed us to infer the processing rate independently for each time period. Table 3 summarizes the arrival counts (per two-hour interval), boarding delays, and estimated processing rates. On average, there are 26.64 arrivals every two hours, the mean boarding delay is 17.88 hours, and the average processing rate is 0.06.

Statistic	Arrival Count (per 2 hours)	Boarding Delay (hours)	Estimated Processing Rate
Count	11700	65582	2925
Mean	26.64	17.88	0.0611
Standard deviation	13.07	15.42	0.0390
Min	1	0.02	0.0000
25%ile	15	5.08	0.0282
50%ile	27	14.78	0.0593
75%ile	37	25.12	0.0808
Max	67	155.03	0.3020

Table 3 Summary statistics

C.3. Additional Results of Case Study

In Section 6.3, we simulate the system under hospital's current policy with thresholds $\bar{b}$. Figure 12 presents additional simulation results for a decision window size (in hours) of $\delta \in \{2, 4, 8, 12\}$. Average waiting time rises while the fraction of rejected transfer requests declines, reflecting the congestion–access trade-off.

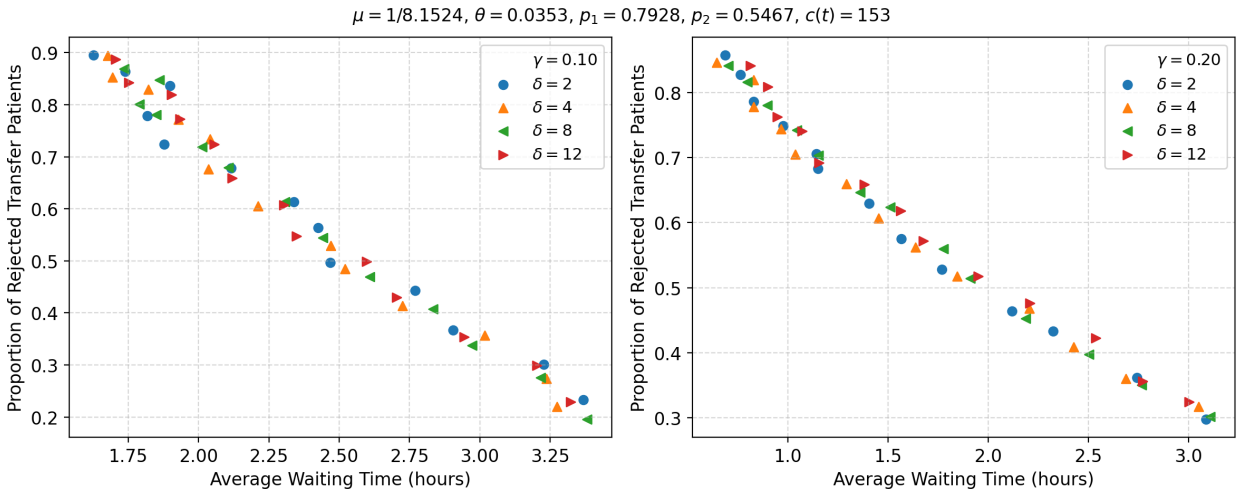

Figure 12 Additional simulation results under hospital's current policy

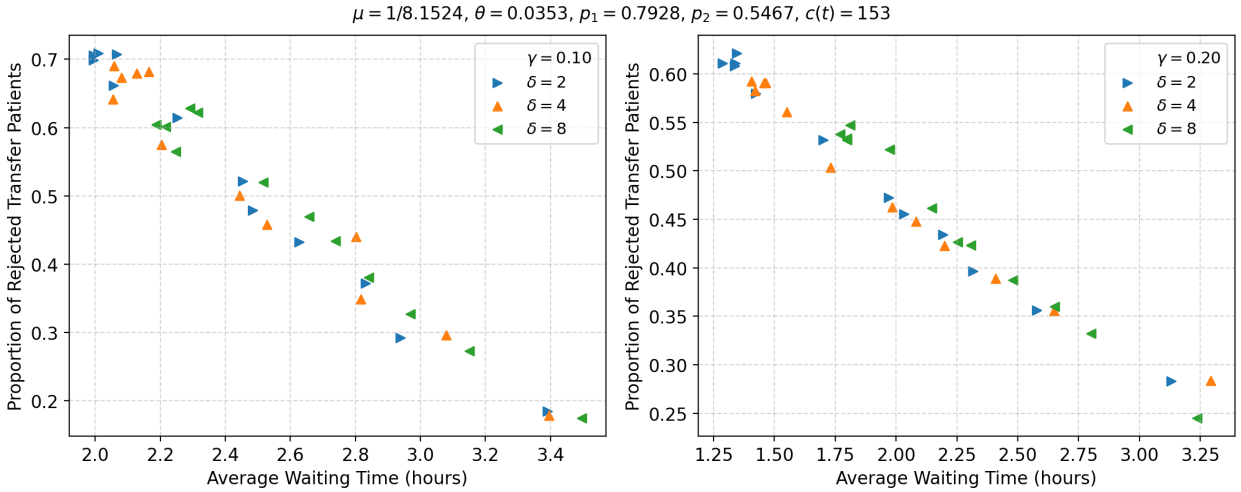

Figure 13 Trade-off between transfer rejection and waiting time under look-ahead policy

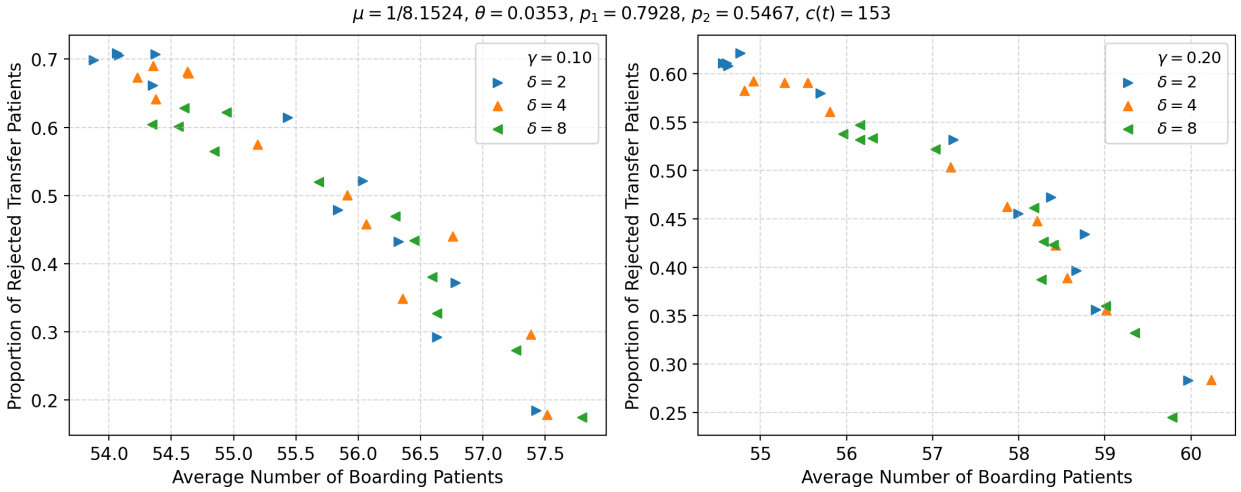

Figure 14 Trade-off between boarding census and waiting time under look-ahead policy

In Section 6.3, we simulate the ED under the adjusted Look-ahead Policy in (12) and Figure 13 presents additional simulation results for a decision window size of $\delta \in \{2, 4, 8\}$. As the ratio of blocking cost to holding cost l/h increases, the system rejects fewer transfer patients, but the average patient waiting time becomes longer. In addition, we report the average hourly boarding census. Specifically, Figure 14 shows the trade-off between the average hourly number of boarding patients during the study period and the rejection of transfer patients for various values of the blocking-cost-to-holding-cost ratio (l/h). As expected, in general, as the average boarding census increases with l/h, the system rejects less transfer patients.

Appendix D: Additional Numerical Experiments

This section reports additional numerical experiments that supplement the results presented in Section 5.

Benchmark policy performance. Figure 15 presents the trade-off curves of benchmark policies with $\eta \in \{0, 0.5, 1\}$ showing that, when Γ is appropriately tuned, the performance of the benchmark policy is largely insensitive to the choice of η.

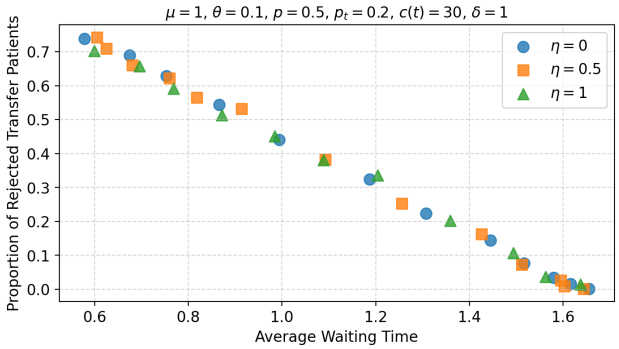

Figure 15 Trade-off under benchmark policy with different η

Time-varying processing rate $\nu(t)$. We next consider a time-varying processing rate $\nu(t)$ with a congestion surge, defined as

$$\nu(t) = \begin{cases} 0.7, & 0 \le t \le 50, \\ 0.3, & 50 < t \le 100, \\ 0.7, & 100 < t \le T, \\ 0, & \text{otherwise.} \end{cases}$$

Figure 16 illustrates the resulting profile of the processing rate $\nu(t)$. The arrival rate is set to $\lambda(t) = 15$ throughout the horizon.

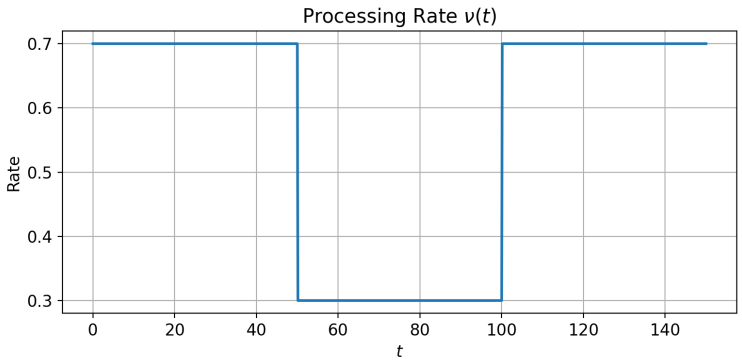

Figure 16 Processing rate $\nu(t)$.

To compare the performance of the benchmark and look-ahead policies, Figure 17 depicts the trade-off between patient waiting time and the proportion of rejected transfer patients. Across both panels (corresponding to review intervals $\delta = 1$ and $\delta = 5$) the look-ahead policy consistently dominates the benchmark by achieving lower rejection rates for comparable or shorter waiting times. These results further highlight the value of incorporating predictive information into proactive admission control decisions.

Prediction errors. In Section 5.2, Scenarios I–V examine all combinations of a single congestion shock, two congestion shocks, $\sigma \in \{1, 2\}$, and $\delta \in \{1, 5\}$. Complete results for all configurations

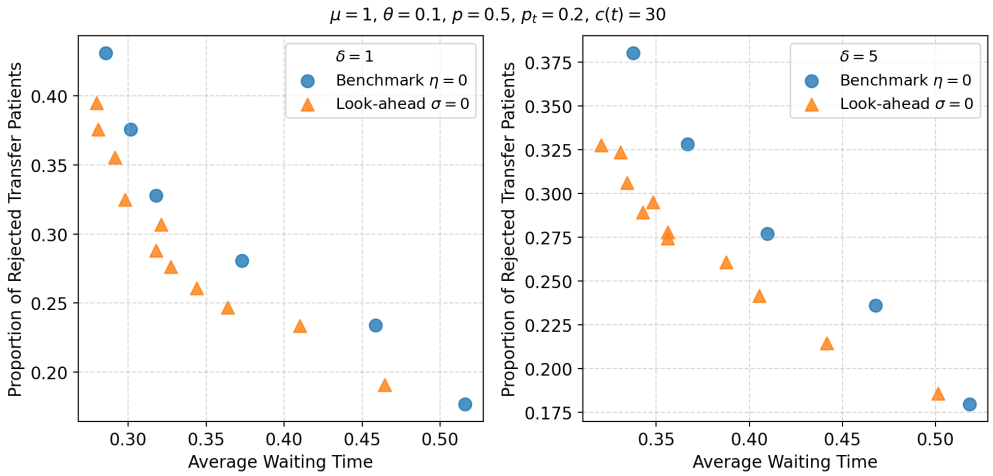

Figure 17 Comparison between look-ahead v.s. benchmark under surge processing rate.

are available at `https://tinyurl.com/e6m37vw2`. The qualitative insights reported in Section 5.2 are consistent across all experiments. We provide some additional examples below.

Figure 18 shows that, under a single congestion shock in Scenario I, policy performance is largely insensitive to the choice of σ for both $\delta \in \{1, 5\}$. In these cases, the look-ahead policy continues to dominate the benchmark. This indicates robustness to moderate levels of unbiased noise. In contrast, Figure 19 shows a degradation in performance under Scenario IV with negative α and a single congestion shock. In this setting, the magnitude of the surge is underestimated, and such biased prediction errors reduce the effectiveness of the proposed policy.

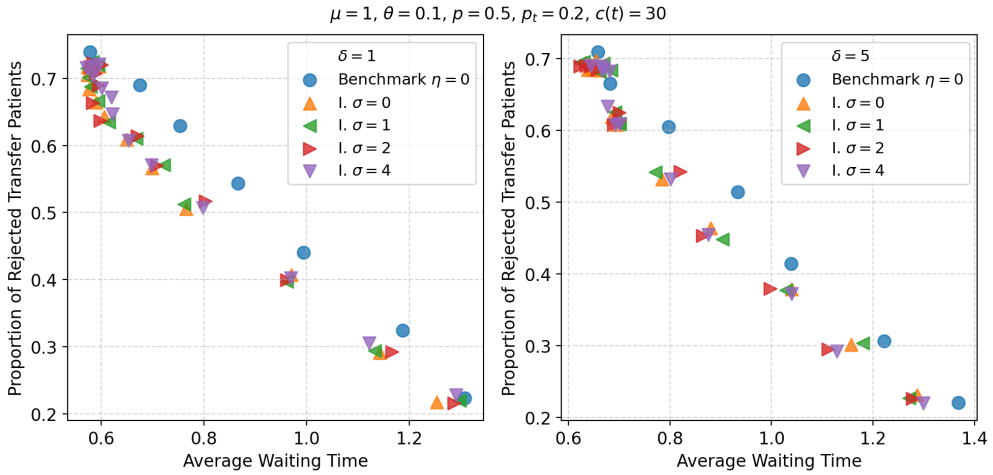

Figure 18 Trade-off comparison between benchmark v.s. look-ahead with unbiased prediction errors

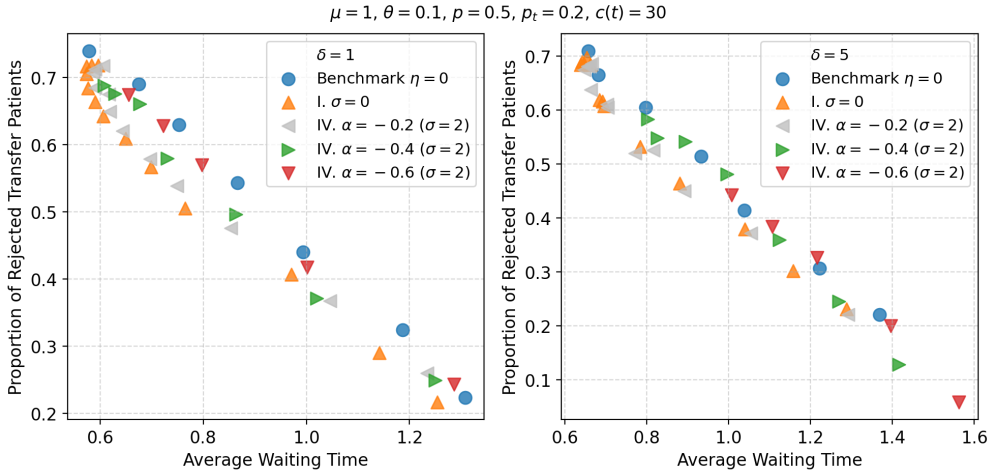

Figure 19 Degradation in performance with biased prediction errors