Using Future Information to Reduce Waiting Times in the Emergency Department via Diversion

Kuang Xu
Stanford Graduate School of Business

kuangxu@stanford.edu

Carri W. Chan
Columbia Business School

cwchan@columbia.edu

This version: October 17, 2015

The development of predictive models in healthcare settings has been growing; one such area is the prediction of patient arrivals to the Emergency Department (ED). The general premise behind these works is that such models may be used to help manage an ED which consistently faces high congestion. In this work, we propose a class of *proactive* policies which utilizes future information of potential patient arrivals to effectively manage admissions into an ED while reducing waiting times for patients who are eventually treated. Instead of the standard strategy of waiting for queues to build before diverting patients, the proposed policy utilizes the predictions to identify when congestion is going to increase and proactively diverts patients before things get 'too bad'. We demonstrate that the proposed policy provides delay improvements over standard policies used in practice. We also consider the impact of errors in the information provided by the predictive models and find that even with noisy predictions, our proposed policies can still outperform (achieving shorter delays while serving the same number of patients) standard diversion policies. If the quality of the predictive model is insufficient, then it is better to ignore the future information and simply rely on real-time, current information for the basis of decision making. Using simulation, we find that our proposed policy can reduce delays by up to 15%.

Key words: Healthcare, queueing, Emergency Departments, predictive models

1. Introduction

Overcrowding in the emergency department (ED) is undesirable as it creates access issues and leads to delays in care. Yet, there is increasing evidence that overcrowding and its subsequent delays frequently occur (Committee on the Future of Emergency Care in the United States 2007, Burt and Schappert 2004). Indeed, 47% of all hospitals in the United States report their ED is at, or even over, capacity (American Hospital Association 2010). In this work, we present an approach which utilizes demand predictions to manage overcrowding in a more effective manner than current strategies used in practice.

There have been many solution approaches which have been suggested to address this overcrowding problem. Some hospitals have resorted to increasing bed capacity to deal with growing demand (Japsen 2003) or using queueing theory to improve staffing decisions (Green et al. 2006). Other approaches have been to encourage and educate patients when it is inappropriate to visit the ED and perhaps more useful to visit their primary care physicians (PCPs) (McCusker and Verdon 2006, Riegel et al. 2002). Another approach used to reduce arrival rates when the ED is overcrowded is ambulance diversion. Crowding in the ED increases the amount of time a hospital spends on diversion (Kolker 2008). Sometimes this crowding is due to congestion in inpatient units (Allon et al. 2013). Even with the effective implementation of these preceding strategies, random variations can still result in periods of overcrowding.

Within the Operations Management community, there has been an extensive body of work examining congestion in the ED. Some of this work has focused on understanding the dynamics of an ED under congestion. For example, Batt and Terwiesch (2012) considers how ED physicians modify the tests they order for patients depending on the number of patients in the ED. Batt and Terwiesch (2015) and Bolandifar et al. (2013) empirically examine how congestion increases patient likelihood of leaving the ED without being seen. Another approach is to use stochastic models to examine how patients should be managed upon arrival to the ED. Saghafian et al. (2012) considers streaming patients based on whether they are likely to be admitted to the hospital or discharged home, while Saghafian et al. (2014) consider utilizing information regarding the amount of ED resources each patient will require in prioritizing patients. Helm et al. (2011) considers admission control to inpatient units and the impact of introducing an expedited patient care queue on reducing ED congestion. Similar to our approach, Dobson et al. (2013) uses a queueing model to provide insight into the management of ED patients. The authors consider how to prioritize new patients versus existing patients in the presence of 'interruptions', while we consider how to determine whether a new patient should even be admitted/treated at the ED versus going somewhere else.

In this work, we consider how predictive modeling can be used to reduce congestion in the ED by leveraging future information when making admission decisions. This future information can potentially be used to effectively reduce arrival rates to the ED in a manner which can substantially reduce the waiting times of those who are actually treated at the ED. We propose and evaluate an algorithm for using predictions of future ED arrivals to make decisions about 'diverting' patients. In this work, we broadly define patient diversion to capture sending patients to various other care options such as diverting ambulances to different hospitals as well as sending low-acuity patients to

urgent care facilities or encouraging PCP appointments. We find that even *noisy* future information can be useful to reduce delays.

There has been a growth in the development of predictive modeling in healthcare. It has been well-documented that arrival patterns to the ED exhibit seasonal patterns. For instance, Green et al. (2006) considers how to modify staffing decisions based on known patterns in arrival rates to the ED. By using a point-wise stationary approximation and utilizing the fact that the majority of patient arrivals occur in the middle of the day, the authors were able to adjust staffing hours in order to reduce waiting times and, subsequently the number of patients who left without being seen. Beyond time-varying arrival rates, predictive models have become much more nuanced and accurate. For instance, Tandberg and Qualls (1994), Rotstein et al. (1997), Jones et al. (2009), Sun et al. (2009) develop predictive models based on time-series analysis to predict emergency department workload. Schweigler et al. (2009), McCarthy et al. (2008), Jones et al. (2002) also examine prediction of ED visits, while Wargon et al. (2009) provides a nice overview. Note that, instead of forecasting just the mean arrival rate for a future time interval, many of these models are capable of making accurate predictions of the **arrival counts**, on a daily (Sun et al. 2009) or even hourly basis (Tandberg and Qualls 1994). The proactive admission policies studied in this paper use the same type of *arrival count predictions*.

A primary motivation in developing these predictive models has been to guide operational decision making, such as 'staff roster and resource planning' (Sun et al. 2009) or 'decisions related to on-call staffing' (Chase et al. 2012). However, while there has been substantial attention paid to developing such predictive models, there has been limited work demonstrating how they can best be utilized to improve system performance. In this work, we take an important first step towards this goal and propose a methodology to consider how predictive models of patient arrival counts could be used to make operational decisions to improve quality of care.

Note that this work bears some similarities to Peck et al. (2012, 2013), which develops predictive models of inpatient unit admissions from the ED and uses them to examine how operational changes–such as prioritizing hospital discharges before noon–can improve flow measures. Our work is differentiated in 2 key ways: 1) we examine different flows of patients, i.e. arrivals to the ED, rather than the discharge from the ED and admission to inpatient units and 2) we leverage a queueing theoretic model and derive analytical results to provide insights into operational decision making, specifically related to admission control and use simulation to verify these insights. The above-mentioned papers primarily rely on simulation models and do not provide analytical results.

Queueing Admission Control with Future Information. Our model is related to the body of literature on Markov queueing admission control, where the decision maker makes dynamic admission decisions while optimizing certain performance objectives. In contrast to our setting, most work in this literature focuses on an *online* problem where future information is not taken into account (cf. Stidham (1985, 2002), and references therein). Our work is also broadly connected to a growing body of work which considers how to use predictive modeling in scheduling, e.g., for satellites (Carr and Hajek 1993), loss systems (Nawijin 1990), and call centers (Gans et al. 2015), which focus on models very different from ours.

Most related to our work is Spencer et al. (2014), which considers strategic 'redirection' of arrivals to an overloaded M/M/1 queue, and demonstrates that it is possible to significantly reduce delay in the heavy-traffic limit by utilizing future information. However, the results from Spencer et al. (2014) fall short in several important aspects in terms of applicability to practical scenarios. In particular, the policy of Spencer et al. (2014) yields substantial delay improvement only when the system is in heavy-traffic, and the performance guarantees apply only when the future information is noiseless; neither the condition of heavy-traffic limit nor that of noiseless future information is likely to be satisfied across the board in most practical systems, including the ED context. In the current work, we generalize beyond the initial insights from Spencer et al. (2014) and propose a family of proactive admission policies that provably outperforms an optimal online policy in any overloaded system, without the need of the heavy-traffic assumption. Furthermore, we investigate the performance of the proposed proactive policies when the future information is noisy, and provide exact performance characterizations for a certain model of prediction noise. While our model and analysis is fairly general and can provide insight into various service settings, our primary motivation is the ED setting where substantial delays can have serious implications for patient care, there has been significant attention on developing predictions of arrival counts, and there may be outside care options (e.g., other hospitals, urgent care facilities, or primary care physicians) to which patients can be 'diverted'.

Our **main contributions** can be summarized as follows:

1. We propose a family of *proactive admission policies*, which, given sufficient future information, delivers superior performance over an optimal online policy at all traffic intensities in the overloaded regime (Theorem 1). To the best of our knowledge, this is the first prediction-guided diversion policy that provably outperforms the optimal online policy in non-heavy traffic regimes.

2. Under a certain 'no-show' model of prediction noise, we quantify the amount of noise tolerance in the predictions of patient arrivals such that our proposed methodology still provides improved

delay guarantees (Theorem 2). We also provide quantifiable guarantees on system performance when only a limited amount of future information is available (Theorem 3).

3. We use simulation to explore the potential reductions in patient delays when utilizing the insights of our analysis to make admission decisions in the Emergency Department. In particular, we demonstrate that the proposed approach can serve the same number of patients as current policies, while, at the same time, reducing the waiting time of patients by up to 15%.

2. Model of the Emergency Department

We describe in this section the ED setting and the subsequent queueing model and decision problem that will form the basis of our investigation. EDs are very complex systems, so while our main queueing model cannot incorporate everything, it captures several key characteristics of the congestion dynamics and provides crucial insights and policy guidelines. In Section 4, we relax some of our modeling assumptions in a simulation model of the ED setting.

2.1. The Emergency Department Setting

Patients can arrive to the ED via ambulance or as a walk-in. New nonurgent walk-in patients can be *diverted* from the ED by encouraging them to go to their Primary Care Physician or an Urgent Care facility (Hoot and Aronsky 2008). Ambulance patients can be *diverted* from the ED via a more formal procedure of 'ambulance diversion' (e.g., Deo and Gurvich (2011)). We use the term *diversion* in the broadest sense to encapsulate both dynamics. While many factors can impact a hospital's decision to divert ambulances or encourage nonurgent patients to receive care elsewhere, congestion in the ED is a one of the main drivers determining ambulance diversion (e.g., Deo and Gurvich (2011), Allon et al. (2013)). As such, threshold policies will serve as a benchmark to which we will compare our proposed proactive diversion policy.

Upon arrival to the ED, a patient is assigned an emergency severity index (ESI) from 1 through 5, with 1 indicating the highest severity. Our analysis will start by focusing on the case of a single patient class. We then discuss extensions to settings with heterogeneous patients and explore the performance of the proposed policy via simulation. Following triage, patients wait until they are taken into an examination room and assigned a bed. While in the examination room, the patient may interact with a physician and nurses, have blood drawn, be taken for various tests, and/or wait between any of the occurrences of these events. Finally, the patient will leave the ED: either being discharged home or admitted to an inpatient unit. In the context of our study, we consider beds as the servers and we define a patient's 'service time' as the time from first treatment to ED discharge, and 'wait time' as the time the patient spends in the waiting area until first treatment.

Though recent work suggests that doctors and nurses may alter behavior depending on the load in the ED (Batt and Terwiesch 2012), we will assume this service time is *exogenous* in order to focus on the impact of diversion on congestion. Moreover, our assumption implies that beds are the bottleneck resource (e.g., Guarisco and Samuelson (2011)). If some other resource (e.g., physicians or testing facilities) were the bottleneck, some work would need to be done to translate our findings to the particulars of the specific scenario.

2.2. Predicting Arrivals

In this work, we consider models which accurately predict the **number** of patient arrivals to the ED. Figure 1 depicts the predicted and realized daily arrival counts to an ED as given in Sun et al. (2009). No predictive model is perfect and their predictive accuracy is often captured by statistics such as the Mean Absolute Percentage Error (MAPE), which ranges from 4.8%-16.9% for daily arrival counts in Sun et al. (2009), and/or the coefficient of determination (R^2), which ranges from 17.7%-42.0% for hourly counts in (Tandberg and Qualls 1994). The proactive policies presented in this paper will similarly utilize noisy predictions of arrival counts.

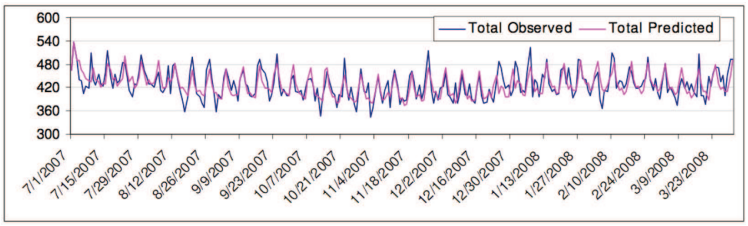


Figure 1 **Predicted versus actual daily arrivals to an emergency department, Sun et al. (2009).**

2.3. Model

Overview of Queueing Model and System Parameters. We will model the waiting area at the Emergency Department as a single, uncapacitated queue, illustrated in Figure 2. The system receives an arrival stream of *homogeneous* jobs (patients) at rate $\tilde{\lambda}$ and is equipped with a total service rate of $\tilde{\mu}$. We will assume that $\tilde{\lambda} > \tilde{\mu}$, in which case we say that the system is in the *overloaded regime*. To stablize the system, the manager is allowed to *divert* incoming jobs up to an average rate of $\tilde{p}$, where $\tilde{p} > \tilde{\lambda} - \tilde{\mu}$, and *admit* the jobs that are not diverted, with the objective of minimizing the average waiting time experienced by all admitted jobs. In practice, the parameters $\tilde{\lambda}$ and $\tilde{\mu}$ can be thought of as system primitives, which can be estimated with reasonable accuracy from historical data, while the diversion rate $\tilde{p}$ can be chosen as a design parameter.

For simplicity of notation, and without loss of generality, we shall normalize $\tilde{\lambda}$, $\tilde{\mu}$ and $\tilde{p}$ by a constant factor of $1/(\tilde{\mu}+\tilde{p})$, to λ, μ and p, respectively, so that $\mu + p = 1$. Equivalently, we have

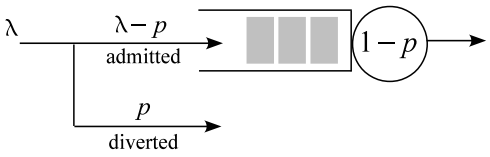


Figure 2 **An illustration of the basic queueing model, where a fraction of the incoming arrivals can be diverted, up to a fixed rate, p.**

that the server runs at rate $1-p$. Under this normalization, the system is fully parameterized by the arrival rate, λ, and diversion rate, p, both in the interval $(0,1)$. We will assume that the value of p is fixed as a constant. The overloaded regime corresponds to having $\lambda \in (1-p, 1)$. To ensure that the admission control problem is non-trivial, we will also assume that $\lambda > p$; otherwise, all arrivals can be diverted. Finally, we shall refer to the limit where $\lambda \to 1$ as the *heavy-traffic regime*, because the resulting arrival rate after diversion, $\lambda - p$, approaches the total system capacity, $1-p$. For the remainder of this paper, we will focus on the non-trivial overloaded regime, with $\lambda \in (\max\{p, 1-p\}, 1)$.

Stochastic Primitives. We will model both the arrival and service processes by Poisson processes in the following ways. Let $\{A(t)\}_{t\in\mathbb{R}_+}$ be the *counting process for arrivals*, where $A(t)$ equals the total number of arrivals to the system by time t. We will assume A to be a Poisson process of rate λ. Similarly, let $\{S(t)\}_{t\in\mathbb{R}_+}$ be the *counting process for service tokens*, where $S(t)$ equals the total number of service tokens produced by time t. We will assume S to be a Poisson process of rate $1-p$, and each of its jumps corresponds to the generation of a service token.

All currently unprocessed jobs are stored in the queue, whose length at time t we denote by $Q(t)$. If there is a jump in A at time t, corresponding to an arriving job, the value of $Q(t)$ is increased by 1. Similarly, if there is a jump in S at time t, corresponding to the generation of a service token, the value of $Q(t)$ is reduced by 1 if and only if $Q(t) > 0$.

Diversion Decisions. Upon arriving to the system, each job is either immediately *admitted* to the queue, where it will wait until it is processed, or *diverted* and leaves the system. A *policy* decides whether an incoming job is admitted or diverted. Denote by $N_\pi(t)$ the number of diversions made by policy π by time t, we define the *average diversion rate* of π as the quantity $\limsup_{t\to\infty} \frac{1}{t}\mathbb{E}(N_\pi(t))$. A policy is considered *feasible* if the resulting average diversion rate does not exceed p. The objective of the decision maker is to choose a feasible policy that minimizes the *average waiting time* in queue experienced by the admitted jobs. We say that a feasible policy, π, is *optimal* among a family of feasible policies, Π, if π achieves the minimum average waiting time among all policies in Π.

Recall that in our ED context, a diversion in our model can correspond to asking a patient with a low acuity level to seek treatment elsewhere–such as at an urgent care facility, an ambulance

diversion to a different facility, or an internal diversion to a different department or medical resource within the same hospital, depending on the specific scenario. In reality, the diversion may cause the demand to increase elsewhere in the system. However, for the purpose of this paper, we will focus on understanding the effect of diversion *locally*, by assuming that the remote facility to which the patients are diverted is relatively congestion-free, and hence the effect of waiting on a diverted patient is negligible. Nevertheless, diversions are still *costly*, both in terms of the direct operational expenses of physical transfers, as well as the additional risk to treatment outcomes for sending a patient to a location with potentially less suitable medical resources than what he or she would have gotten in the original ED. This is captured by the upper bound on the average rate of diversion, p, which can be chosen by the decision maker as a design parameter.

Future Information. The amount of prediction or future information the decision maker has is characterized by a lookahead window: at any time t, the decision maker is endowed with a *lookahead window* of length w, which comprises of the (possibly noisy) *prediction* of the arrival and service token processes, i.e., A and S, in the time interval $[t, t+w)$. In particular, the case of $w=0$ corresponds to the *online* setting, where no future information is available, and all diversion decisions have to be made solely based on the current state of the system. At the other extreme, the case of $w=\infty$ corresponds to an idealized case of *infinite lookahead*, where the time of all future arrivals and service tokens have been revealed. In reality, the future may be predictable only within a small time window, and hence we will be mostly interested in the case where w is finite. As our model of future information provides predictions for both the arrival and service token processes, but the focus of the ED literature is on predicting arrival counts, we will relax the assumption regarding knowledge of the service token process in our simulations and find that the proposed policy still outperforms the benchmark.

Implications of Service Tokens. The use of a service token process corresponds to the case where the jobs' service times are induced by an *exogenous process*, e.g., randomness in the speed of the server, and are decoupled from the jobs' identities. The use of service token processes has several benefits. Firstly, as a result of the decoupling between the job's identity and its service time, the use of service tokens also ensures that the resulting queue length process is *insensitive* to the *service priorities* adopted at the server (i.e., which jobs to serve first). In particular, because the generation of the service tokens does not depend on the identities of the jobs currently in the queue, it is not difficult to show that, fixing the admission control policy and the sample paths for the arrival and service token processes, the sample path of the queue length process remains the same under any work-conserving service rule adopted by the server. Additionally, this invariance to

service priorities holds because our objective is to minimize average waiting time (Deo and Gurvich 2011). Next, while the original admission control problem is formulated for a queue with a *single* server, the setup allows for an easy heuristic approach when considering the case of *multiple* servers. In particular, when the system is equipped with k servers, once could approximate the system as one with a Poisson service token process of rate that is k times the original. While this approach does not perfectly model an $M/M/k$ setting, it provides a rough estimate of system dynamics in the multiserver setting. We expect the proposed policy, with analytic guarantees for $k = 1$, to perform well in the multiserver setting and, as will be seen in Section 4, simulation results concur.

In an online setting (without future information), it is not difficult to show that the queue length process in the system with service tokens is equivalent to that of the total number of jobs in system for an $M/M/1$ queue, where the jobs' service times are i.i.d, because both evolve according to a birth-death process with birth rate λ and death rate $1-p$. However, this equivalence relation no longer holds when future information is involved. For instance, the service token setup is not equivalent to the case where the incoming jobs are associated with their own service times which are known in advance, and it should only serve as an approximation for this setting. Thus, while we leverage the token process assumption for our analytic results, we relax this assumption in our simulation model and find that we still achieve substantial performance gains.

3. Proactive Diversion Policies and Analytical Results

We describe in this section the family of proactive diversion policies, as well as our main analytical results concerning their performance in terms of delay and tolerance to prediction noise. All proofs are given in the Appendix.

We first provide some intuition on why a proactive policy, which acts upon the knowledge of future events, may outperform an online policy. It is well known that, in the online setting, the admission control problem considered in this paper admits an optimal policy that is of a *threshold* form (cf. Stidham (1985, 2002), Spencer et al. (2014)), where, for some fixed threshold L, an arrival at time t is diverted if and only if $Q(t) = L$. We shall denote by $TH(L)$ this *threshold diversion policy* with threshold L. It is not difficult to show that as the threshold L increases, the diversion rate induced by $TH(L)$ decreases and the resulting expected queue length in steady-state increases (cf. Eq. (5.10) of Spencer et al. (2014)). Therefore, the optimal value of L corresponds to choosing the smallest L for which the diversion rate is no more than p, as seen in Spencer et al. (2014). For any $p \in (0,1)$ and $\lambda \in (\max\{p, 1-p\}, 1)$, this leads to:

$$L = L(p,\lambda) = \log_{\frac{\lambda}{1-p}} \frac{p}{1-\lambda} - 1. \tag{1}$$

Note that to avoid excessive use of ceilings and floors in our notation, we will assume that $\log_{\frac{\lambda}{1-p}} \frac{p}{1-\lambda}$ is an integer. The resulting online-optimal expected queue length is then:

$$\mathbb{E}(Q) = \frac{p}{\lambda - (1-p)} L(p, \lambda) + \frac{\lambda(1-\lambda) - p(1-p)}{(1-\lambda-p)^2}. \tag{2}$$

The threshold policy $TH(L)$, while being optimal for the online setting, suffers from a drawback that is in some way inevitable for all online policies: it is allowed to make diversions only when the queue is *large*, i.e., at length L. Specifically, imagine a scenario where the system is about to encounter a 'bursty episode', during which there are relatively more arrivals than service completions. Not having access to information about the future, a threshold policy will have to wait until the queue builds up to length L before it starts to make diversions, after which point the long queue is bound to cause large delays for subsequent arrivals. In contrast, knowing the onset of a bursty episode beforehand, a proactive policy can make diversions *earlier*, and potentially prevent the queue from building up to length L in the first place. Indeed, our result (Theorem 1) shows that such a proactive policy can achieve a substantially smaller queue length than that of the threshold policy (Eq. (2)), at the same diversion rate.

It remains, however, to precisely define what it means to 'divert early' in a proactive policy. To this end, we will make use of an indicator that separates the arrivals that appear at the beginning of a 'bursty episode', whose diversion could significantly reduce the waiting time experienced by subsequent arrivals, from those arrivals that appear later, whose diversion will have relatively less impact on others and should hence be admitted. This indicator, which we refer to as being w-blocking, will be defined in Definition 1, and it will be used as a main input to our proactive diversion policies, in Definition 2.

3.1. Proactive Policies with Thresholds

Denote by $\{Q^0(t)\}_{t\in\mathbb{R}_+}$ the *baseline queue length process* generated by A and S, where $Q^0(t)$ is the queue length at time t assuming no diversions are made, i.e.,

$$Q^0(t) = \sup_{0\le s\le t} (A(t) - A(s)) - (S(t) - S(s)). \tag{3}$$

Note that because we are operating in the overloaded regime, $Q^0(t) \to \infty$ as $t \to \infty$.

Definition 1 (w-Blocking Arrivals) *A job that arrives at time t is* w-blocking *if*

$$\min_{0\le s<w} Q^0(t+s) \ge Q^0(t^+). \tag{4}$$

where $f(x^+)$ denotes the right limit of f at x.

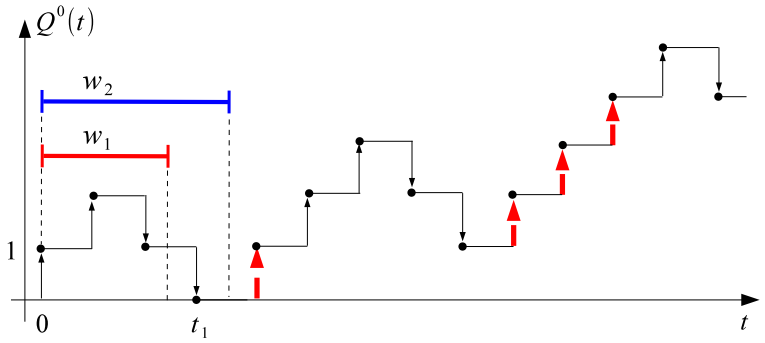


Figure 3 **An example sample path of the baseline queue length process, Q^0. The bold arrows correspond to the ∞-blocking arrivals. Note that if the lookahead window is too short, with $w = w_1 < t_1$, then the arrival at $t = 0$ would be an w-blocking arrival, because it is not until time t_1 that the baseline queue length process 'goes below' $Q^0(0) = 1$, an event that cannot be foreseen within the lookahead window at $t = 0$. This is not be the case with a sufficiently long lookhead window, e.g., $w = w_2 > t_1$.**

Note that whether an arrival at time t is w-blocking is fully determined by the realizations of A and S in the interval $[t, t+w)$. In words, an arrival at time t is w-blocking if the baseline queue length process will *not* return to its current level at time t within the next w units of time. As an alternative interpretation, assuming the baseline queue length process is zero at time t, then w-blocking corresponds to the busy period associated with the arrival in the baseline queue length process being longer than w units of time (See Figure 3 for an example). These w-blocking arrivals correspond to the jobs that arrive at the *beginning* of a 'bursty episode', during which there are more arrivals than service tokens. Intuitively, these jobs, if admitted, tend to delay all *subsequent* arrivals during the bursty episode (hence the nomenclature 'blocking'), and their diversions are beneficial for reducing the overall delay.

This notion of w-blocking arrivals was first introduced by Spencer et al. (2014) in the design of their NOB admission policy, which amounts to diverting all, and only, w-blocking arrivals. We define a more general family of proactive diversion policies, which utilizes the notion of w-blocking:

Definition 2 (w-Proactive Policies with Thresholds) *Fix $s \in \mathbb{Z}_+$ and $l \in \mathbb{Z}_+ \cup \{\infty\}$, with $s < l$, and let $Q(t)$ be the queue length at time t. Under a w-proactive diversion policy with thresholds (s,l), denoted by $PA_w(s,l)$, an arrival at time t is diverted if and only if,*

1. *$Q(t) = l$, or*
2. *$Q(t) \geq s$, and the arrival is w-blocking.*

We make some simple observations of the w-proactive policies.

1. First, note that an arriving job that is w-blocking will always be diverted whenever the queue length at the time of its arrival lies within the range of $[s,l]$. Such diversions of w-blocking arrivals

constitute the 'proactive' aspect of the policy, which allow it to respond quickly to surges in arrivals in the near future.

2. Second, an arrival will always be diverted if the current queue length is at level l, regardless of whether it is w-blocking. This upper-threshold pushes the proactive policy to be more aggressive when the queue length is excessively long, and it will prove to be critical in helping the proactive policy maintain an advantage in queue length over an optimal online policy (Theorem 1).

3. Finally, no diversion is to be made when the queue length is less than s. The lower-threshold s reduces the rate of diversion by disallowing any diversion when the queue length is too small. Note that by setting s closer to l, the behavior of the proactive policy will become closer to that of an online threshold policy with threshold l. Although the lower-threshold is not essential if there is an abundance of future information (large w) and the predictions are noiseless, the application of a lower-threshold becomes critical as it ensures that the proactive diversion policy can be made *feasible* even under limited and noisy predictions (See Section 4).

The family of $PA_w(s,l)$ policies can also be viewed as a framework that generalizes both the online threshold policy $TH(L)$ and the NOB policy proposed in Spencer et al. (2014). In particular, the $TH(L)$ policy corresponds to the policy $PA_w(L,L)$, where the two thresholds are both equal to L, and the policy in Spencer et al. (2014) corresponds to the policy $PA_w(0,\infty)$, where neither the lower nor upper thresholds are applied and the policy diverts only the w-blocking arrivals.

We highlight two main benefits of such a generalization, compared to the previous approaches. First, the $PA_w(s,l)$ policy provides the flexibility that would allow a manager to smoothly transition between proactive versus online decision making by simply modifying the values of the thresholds s and l, depending on the amount and quality of future information available, without changing the inner logic of the algorithm. Second, with appropriately chosen threshold values, the $PA_w(s,l)$ policy is able to strictly outperform both the optimal online threshold policy, and the NOB policy of Spencer et al. (2014), at all traffic intensities in the overloaded regime, given sufficient future information (cf. Theorem 1 and Figure 4). To our knowledge, this is the first prediction-guided admission policy that provably outperforms the optimal online policy at all traffic intensities in an overloaded system. (Note that the policy of Spencer et al. (2014) is only guaranteed to outperform the online policy in heavy traffic as $\lambda \to 1$.)

One feature of this proposed set of proactive diversion policies is that two arrivals may experience a different admission decision (one is admitted while the other is diverted) even if the system appears to be the same, i.e. having the same queue length, for both arrivals. That said, the diversion

policy is completely agnostic to patient information beyond the state of the system upon arrival– just as the current, online threshold-based policies are. In this sense, all patients are fairly treated in the same manner under the proposed proactive policies. In both the online and proactive setting, if a patient arrives at an 'inopportune time', he will be diverted. Determining whether the current epoch is an 'inopportune time' now depends on future information and the current queue length, whereas in the online setting, it only depends on the current queue length.

Systems with Priority Arrivals. While we have thus far focused on a homogeneous system where the incoming jobs are treated as identical, our methodology can be extended to incorporate *service priority* as well. There are two main types of priority in our setting:

1. The order in which the admitted patients are served may not be first-come-first-serve.
2. There is a subset of the arrivals (e.g., ambulance arrivals) that must be admitted and hence *cannot* be considered for diversion.

For the first type of priority, we note that the service token model we adopted already implies that the average delay experienced by the admitted jobs are *insensitive* to the order in which they are served, so long as the service policy is work-conserving. One could address the second type of priority by using a natural extension of the w-proactive policy, where the impact on the service availability induced by the set of prioritized arrivals is incorporated into the calculations of the baseline queue length process (See Appendix B.2 for more details). In our simulations, we will consider both forms of patient heterogeneity and find the proposed policy still outperforms the benchmark.

3.2. Delay Improvement from Proactive Policies in Moderate Traffic

We present our analytical results in the next three subsections. To build intuition, we will first focus on the case of infinite lookahead (i.e., $w = \infty$). We will then discuss, in Section 3.4, how the insights from our analysis can be extended to the finite-lookahead case (i.e., $w < \infty$). We will also assume that the realizations of service tokens, but not the arrival tokens, can be predicted noiselessly within the lookahead window. Simulations in Section 4 examine the more realistic case where only the mean of the service times is known.

Our first main finding shows that the proposed family of proactive policies is capable of strictly improving upon the delay performance of an online policy in expectation, at all traffic intensities in the overloaded regime.

Theorem 1 *Fix* $p \in (0,1)$ *and* $\lambda \in (\max\{p, 1-p\}, 1)$.

1. *Let π_i be the steady-state probability of $Q=i$ under the $PA_\infty(0,l)$ policy, then*

$$\pi_i = \begin{cases} \frac{1-\beta}{1-\beta^{l+1}}\beta^i, & 0 \le i \le l, \\ 0, & \textit{otherwise}, \end{cases} \tag{5}$$

where $\beta = \frac{1-p}{\lambda}$.

2. *The optimal threshold, l^*, for the $PA_\infty(0,l)$ policy is given by*

$$l^* = L(p,\lambda) = \log_{\frac{\lambda}{1-p}} \frac{p}{1-\lambda} - 1. \tag{6}$$

In particular, l^* coincides with the threshold used in the optimal online policy (cf. Eq. (1)).

3. *Denote by $\mathbb{E}(Q_{PA^*})$ and $\mathbb{E}(Q_{ON})$ the steady-state expected queue lengths under the $PA_\infty(0,l^*)$ policy and an optimal online policy, respectively, and by $\mathbb{E}(W_{PA^*})$ and $\mathbb{E}(W_{ON})$ the corresponding expected steady-state waiting times. We have that*

$$\mathbb{E}(Q_{PA^*}) = \frac{1-\lambda}{1-\lambda-p} L(p,\lambda) - \frac{\lambda(1-\lambda)-p(1-p)}{(1-\lambda-p)^2}, \tag{7}$$

and

$$\mathbb{E}(Q_{ON}) - \mathbb{E}(Q_{PA^*}) = \frac{(1-\lambda)+p}{\lambda+(1-p)} L(p,\lambda) + 2\frac{\lambda(1-\lambda)-p(1-p)}{(1-\lambda-p)^2} > 0, \tag{8}$$

for all $\lambda \in (\max\{p, 1-p\}, 1)$. By Little's law, the above equations further imply that

$$\mathbb{E}(W_{PA^*}) = \frac{1}{\lambda-p}\left[\frac{1-\lambda}{1-\lambda-p} L(p,\lambda) - \frac{\lambda(1-\lambda)-p(1-p)}{(1-\lambda-p)^2}\right], \tag{9}$$

and

$$\mathbb{E}(W_{ON}) - \mathbb{E}(W_{PA^*}) = \frac{1}{\lambda-p}\left[\mathbb{E}(Q_{ON}) - \mathbb{E}(Q_{PA^*})\right] > 0, \tag{10}$$

for all $\lambda \in (\max\{p, 1-p\}, 1)$. In other words, the $PA_\infty(0,l^*)$ policy strictly improves upon the optimal online policy at all traffic intensities in the overloaded regime.

While the theorem applies to $w=\infty$, we relax this requirement in Section 3.4.

This result formalizes the intuition discussed at the beginning of Section 3 as to how proactive diversions can be helpful. Figure 4 illustrates the delay improvements of the proactive policy compared to an optimal threshold policy and the NOB policy in Spencer et al. (2014). We see that the gains over the online policy are most substantial when the system is overloaded; that said, the proactive policy is still doing better in moderate traffic. Indeed, it is not difficult to deduce from this result that the average queue length induced by the $PA_\infty(0,l)$ policy monotonically decreases as l decreases. It has been shown in Spencer et al. (2014) that the $PA_\infty(0,\infty)$ policy achieves the optimal average queue length in the heavy-traffic regime of $\lambda \to 1$, among all diversion policies

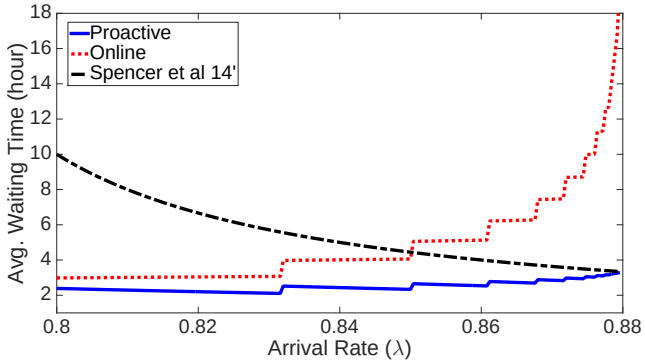


Figure 4 **Comparison of the average waiting time under the proactive policy $PA_\infty(0, L(p,\lambda))$ (Eq. (7)), the optimal online policy $TH(L)$ (Eq. (2)), and the NOB policy in Spencer et al. (2014), with $p = 0.3$. The proactive policy achieves a better average waiting time for all values of λ in the interval $(\max\{p, 1-p\}, 1)$. The performance of the proactive policy converges to that of the NOB policy as $\lambda \to 1$.**

that utilizes future information. It hence follows that the $PA_\infty(0, l^*)$ policy also admits the delay optimality in the heavy-traffic regime, and its resulting average queue length approaches that of $PA_\infty(0,\infty)$ from *below*, as $\lambda \to 1$, as is illustrated in Figure 4. It is the presence of the upper threshold, l, which partially aligns the proactive policy with the online one, that is essential to guarantee better performance for all traffic intensities in the overloaded regime.

Theorem 1 also provides valuable insights on how to apply the proactive policy $PA_w(s,l)$ in practice. A key conclusion from Theorem 1 is that choosing the upper-threshold l to be equal to that of the optimal online threshold policy, and the lower-threshold s to be zero, yields superior delay performance. In practice, however, the lookahead window is finite ($w < \infty$) and the predictions are noisy, and hence we should not expect the same to hold exactly. Nevertheless, it is reasonable to consider a similar heuristic, by choosing the upper-threshold, l, to be close to that of the online threshold policy, and then find a sufficiently large lower-threshold, s, to ensure feasibility. This heuristic substantially simplifies the search for the optimal choices of the thresholds, and evidence from simulations shows that it is capable of finding (s,l) pairs that are near-optimal (See more details in Section 4.5).

3.3. Noisy Predictions of Arrivals

Our analysis thus far has assumed that the future information which is available is *perfect* in the sense that arrival and service tokens are known exactly. However, it is impractical to assume that predictive models will have this kind of predictive power. At best, the future information is noisy. Hence, we consider a scenario where the arrival observations in the lookahead window are a *noisy* version of the actual realizations. A natural question is whether using such noisy information can still improve delay or if it is better to simply ignore the future information and resort to the online

diversion policies. The main analytical result in this section quantifies the performance impact under a certain noise model where predicted arrivals may not actually be realized in the real system. We also discuss the case where the future information failed to provide any indication that there may be such an arrival, and use simulation in Section 4 to further confirm that our proactive policy performs well under more realistic noise models.

3.3.1. No-show Noise We start by considering the implications of a type of no-show noise– predicted arrivals never appear in the realized process. That is, the prediction algorithm is able to foresee all potential arrivals, while some of them may not be realized in the actual arrival process. More precisely, we consider the following model of noisy predictions: Fix $\lambda \in (\max\{p, 1-p\}, 1)$. Let $\epsilon \in [0,1)$ be a known parameter which specifies the *level of arrival no-shows*. Let $\{A'(t)\}_{t\in\mathbb{R}_+}$ be the *predicted* arrival process so that A' is a Poisson process with rate $\lambda/(1-\epsilon)$. Each arrival in A' belongs to the actual arrival process, A, (corresponding to a 'realized' arrival) with probability $1-\epsilon$; with probability ϵ the predicted arrival is not realized in the actual arrival process and can be considered a 'no-show'. Thus, the actual arrival process, A, consists of a proportion $1-\epsilon$ of the arrivals in A'. Note that by the thinning property of Poisson processes, A is a Poisson process of rate λ.

The no-show noise model with parameter ϵ can be thought of as a special case of *noisy predictive models for arrival counts*, discussed in the Introduction. In particular, it can be associated with a certain predictive model for future arrival counts with a Mean Absolute Percentage Error of $\epsilon/(1-\epsilon)$ or a coefficient of determination (R^2 coefficient) of $(1-2\epsilon)/(1-\epsilon)$. See Appendix B.1 for more details. In practice, the parameter ϵ may be estimated from historical data, by examining the fraction of predicted arrivals that has failed to materialize. In considering this noise setting, we characterize the size of noise parameter ϵ such that a proactive policy maintains the same delay guarantees while remaining feasible (i.e. diverting at most p fraction of *realized* arrivals).

Theorem 2 *Consider the no-show noise model with a level of arrival no-shows $\epsilon \in [0,1)$. If $l < \infty$, then the $PA_\infty(0,l)$ policy is feasible if and only ϵ satisfies*

$$\lambda - (1-\epsilon)(1-p)\frac{1-\beta^l}{1-\beta^{l+1}} \le p, \tag{11}$$

where $\beta = (1-\epsilon)^2\frac{1-p}{\lambda}$, and the resulting steady-state expected queue length and waiting time are given by

$$\mathbb{E}(Q) = \frac{l}{1-\beta^{-(l+1)}} - \frac{\beta(\beta^l-1)}{(\beta-1)(\beta^{l+1}-1)}, \quad \mathbb{E}(W) = \frac{1}{\lambda-p}\left[\frac{l}{1-\beta^{-(l+1)}} - \frac{\beta(\beta^l-1)}{(\beta-1)(\beta^{l+1}-1)}\right]. \tag{12}$$

Note that Theorem 2 also holds in the limit as $l \to \infty$.

This result quantifies the amount of noise tolerance of our delay guarantees. In particular, it provides a basis on which one can determine if a predictive model is 'good enough', or if more work is necessary to improve its predictive power (i.e. reduce noise in the prediction) before it can be used to help manage admission decisions in an effective manner. It also follows from Theorem 2 that the expected queue length in steady-state is monotonically decreasing as ϵ increases (Eq. (12)). This implies that the delay performance of the proactive policy does *not* degenerate in the presence of no-show noise; instead, we pay a price in terms of an increased rate of diversion.

In practice, ϵ can be estimated by in-sample or out-of-sample performance of the predictive model as measured via historical data. The MAPE for the predictive models in Sun et al. (2009) range from 4.8%-16.9%, implying that in practice ϵ may be between $[.05, .20]$. On the other hand, the R^2 for the predictive models in Tandberg and Qualls (1994) range from 17.7%-42%, implying a range of $\epsilon \in [.37, .45]$.

3.3.2. Unpredicted Arrivals We now discuss the implication of our models in the setting where some arrivals cannot be predicted. Thus, this captures both noise factors discussed earlier: some predicted arrivals will not show up while another set of arrivals are not observed by the predicted model. In this case, one can think of the arrival process A as a superposition of two processes, $A(t) = A_p(t) + A_u(t)$, where A_p and A_u correspond to the predicted and unpredicted arrivals, respectively. Assuming the manager knows whether an arrival belongs to A_p or A_u, a simple way to handle this setting is by dividing the service capacities, and our case, the service token process S, into two corresponding portions: $S(t) = S_p(t) + S_u(t)$, whereby we use the process S_p to serve the predicted arrivals A_p using the algorithms discussed in this paper, and use the process S_u to serve the unpredicted arrivals A_u by applying online admission control policies. Note that since each stream is independent and feasible, the feasibility of this approach is guaranteed. Moreover, by the delay improvements for the predictable stream, when the fraction of unpredicted arrivals is relatively small, this split approach is guaranteed to have lower delays than a purely online policy in the heavy-traffic regime. We note that, in some cases, the predicted and unpredicted arrivals may be correlated. It is not clear whether Poisson processes remain valid for modeling such a setting, and it can be an interesting topic for future research.

3.4. From Infinite to Finite Lookahead

We now consider the scenario where the length of the lookahead window, w, is *finite*. Our analysis in this subsection will focus on the case where the prediction in the lookahead window is noiseless.

Still, we expect that by adapting and incorporating the steps from the proof of Theorem 2 it will be possible to establish analogous results for the no-show prediction noise model in Section 3.3.

We will focus on the performance of a $PA_w(0,l)$ policy where $w < \infty$. Decreasing w strictly enlarges the set of w-blocking arrivals because decreasing the value of w makes the proactive policy become more aggressive and divert more jobs. On the positive side, the enlargement of the set of diversions implies that the average queue length under $PA_w(0,l)$ is non-increasing as w decreases. Therefore, the expression on the expected queue length under $PA_\infty(0,l)$ given in Eq. (7) automatically serves as an upper bound for the average queue length when $w < \infty$.

On the negative side, however, the diversion rate of $PA_w(0,l)$ increases as we decrease the value of w, which could lead to over-diversion when w is too small. In order to ensure that $PA_w(0,l)$ is a feasible policy, it is important to quantify the changes in the diversion rate as a function of w. The main result of this subsection provides upper and lower bounds on the diversion rate induced by the $PA_w(0,l)$ policy for all values of $w \in \mathbb{R}_+$. Denote by $F_{a,b}(\cdot)$ the cumulative distribution function for the busy period distribution of an $M/M/1$ queue with arrival rate a and service rate b (cf. Chapter 2, Gross et al. (2013)),

$$F_{a,b}(x) = \int_0^x \frac{1}{s\sqrt{a/b}} e^{(a+b)s} I_1(2s\sqrt{ab})\, ds, \tag{13}$$

where $I_1(\cdot)$ is the modified Bessel function of the first kind of order one, with $I_1(x) = \sum_{k=0}^{\infty} \frac{(x/2)^{2k+1}}{k!\,(k+1)!}$. We have the following characterization of the diversion rate.

Theorem 3 *Fix $w \in \mathbb{R}_+$ and $l \in \mathbb{Z}_+$, and let $\beta = \frac{1-p}{\lambda}$. Denote by $r_{w,l}$ the diversion rate induced by the $PA_w(0,l)$ policy. We have that*

$$r_{w,l} \le \lambda - (1-p)\frac{1-\beta^l}{1-\beta^{l+1}} + (1-p)\left(1 - F_{1-p,\lambda}(w)\right), \tag{14}$$

and

$$r_{w,l} \ge \lambda - (1-p)F_{1-p,\lambda}(w). \tag{15}$$

Theorem 3 provides both a quantitative and qualitative assessment of when the w-proactive policy should work, and when it may fail. On the one hand, for any λ, the last term in the right-hand side of the upper bound (Eq. (14)) converges to 0 as $w \to \infty$; hence, we conclude that the diversion rate of $PA_w(0,l)$ converges to that of $PA_\infty(0,l)$. In other words, the $PA_w(0,l)$ policy is feasible so long as there is sufficient future information, relative to the value of λ.

On the other hand, the lower bound on $r_{w,l}$ in Theorem 3 shows that, as $w \to 0$, the diversion rate of $PA_w(0,l)$ converges to λ,which is strictly greater than the maximum diversion rate, p. Therefore,

additional measures must be taken to reduce the diversion rate, for otherwise the proactive diversion policy is bound to become infeasible when the size of the lookahead window is too small. As was mentioned in Section 3.1, this effect of over-diversion motivates us to incorporate a *lower threshold* in our diversion policy so that no job is diverted when the queue is too small, which corresponds to a $PA_w(s,l)$ policy with $s>0$. With an appropriately chosen lower threshold, our simulations results in Section 4 show that it is possible to maintain the feasibility of the proactive diversion policy, while keeping the average queue length small.

Finally, despite the diversion rate guarantees provided by Theorem 3, when w is small, we can no longer ensure that the resulting delay given by the best $PA_w(0,l)$ is strictly smaller than that of an optimal online policy for all $\lambda \in (1-p,1)$, unlike in Theorem 1. Indeed, it is shown in Xu (2015) that delay cannot be improved by more than a constant factor beyond the optimal online policy, if w is substantially smaller than $\Theta\left(\frac{1}{1-\lambda}\right)$. It remains an open question whether it is possible to find a diversion policy that provably outperforms the online policies at all traffic intensities even when w is small.

4. Simulation Results

We now examine the insights of our model and analysis via a simulation, which captures a number of features present in the ED setting. We will see that the proactive policy with thresholds is able to consistently outperform the online threshold policy under different levels of prediction noise and diversion rates.

System dynamics. We consider an emergency department with 20 beds, corresponding to a medium-sized ED (e.g., Saghafian et al. (2014) simulates a 22-bed ED, and the main ED in Khare et al. (2009) has 23 beds). Each bed is represented by a server in the simulation. We discretize the continuous time into two time scales. We will assume that the queueing dynamics and the diversion decisions operate on the basis of 15-minute time slots, whereas the predictions of arrivals are performed on an hourly basis (i.e., every 4 time slots). This assumption is more stringent than the one used in our theoretical model, where the time scale of the predictive model is the same as the underlying queueing dynamics. However, we believe that this models reality more closely, where it is difficult to make arrival predictions at the finest time granularity (see Tandberg and Qualls (1994) for a model which provides hourly predictions). While diversion decisions can occur on a more continuous time line, it does take time to implement such decisions, so a 15 minute granularity is sufficient for illustrative purposes. More precisely, for each hour, k, we will generate a (noisy) prediction of the total number of arrivals during the hour, $a_{pred}(k)$. Then, to generate the predicted arrival sample path used in the proactive policy, we assign each of the

$a_{pred}(k)$ arrivals uniformly at random to the 4 time slots within the hour. All numerical results are obtained by averaging over 100 runs of simulations, each over a one-year time span and a one-month warm-up. Since we are averaging waiting times over a year, the 95% confidence interval for all of our simulations is tighter than ± 1 second around the reported average waiting time. As such, we do not explicitly report the confidence intervals in our presentation of our simulation results.

Arrivals. We use a Poisson process with time-varying rates to model the arrivals. In particular, the number of arrivals in a time slot is a Poisson random variables, independent from all other slots, whose mean depends only on the hour of the day and the day of the week (i.e., intra-week rate variations are not considered). The hourly arrival rates are obtained from emergency room records in the SEEStat database (SEE-Center (2009)), averaged across the year of 2004. We will express the average arrival rate (over the time-varying rates), $\tilde{\lambda}$, as a multiple of the total service capacity. The value of $\tilde{\lambda}$ is generally greater than 1, corresponding to the overloaded regime, and is initially fixed to be 1.2. To begin, we assume that all arrivals are within the same triage class, and can be subject to diversion. We consider multiple patient types in Section 4.4.

Service times. Differing from the service token assumptions used in our theoretical analysis, which was useful to allow for analysis, we will simulate the more practical scenario where the service times are attached to individual jobs, and that the lengths of the service times are *unobserved*. We assume that the service times are mutually independent and distributed according to a lognormal distribution with a mean of 3 hours, truncated to a maximum of 24 hours (e.g., Batt and Terwiesch (2012, 2015)). Because the actual service time of a job is unobserved, we use the mean service time in its place when generating the baseline queue length process Q^0 which is needed to identify w-blocking jobs.

Noise Model. Unless otherwise specified, we assume that the decision maker is able to make noisy predictions of the number of arrivals in *each hour* within a 24-hour lookahead window. We will model the prediction noise by assuming that the observed number of arrivals during each hour deviates from the true realization by a normally distributed perturbation (cf. Schweigler et al. (2009), Sun et al. (2009)), drawn i.i.d. for each hour. The magnitude of the noise is parameterized by a coefficient, q, in the following fashion. Letting $a(k)$ and $a_{pred}(k)$ be the actual and predicted numbers of arrivals during the kth hour, respectively, then $a_{pred}(k)$ is equal to $a(k) + N(0,1)\sqrt{q\text{Var}(a(k))}$, rounded to the nearest non-negative integer, where $N(0,1)$ denotes a standard normal random variable. In other words, q corresponds to the predictive model's *expected least-squared error* relative to the arrival variance, or, equivalently, the value of $1 - R^2$, where the R^2 is the predictive model's coefficient of determination. Relating back to the no-show noise model in Section 3.3.1,

the parameter q satisfies the following relationship $\epsilon = q/(1+q)$ (cf. Appendix B.1). The case of $q = 0$ corresponds to that of perfect predictions. As $q \to 1$, the variance of the noise approaches that of $a(k)$ itself, which is essentially the same as simply using $\mathbb{E}(a(k))$ as a predictor. Consequently, it is sufficient to restrict our attention to the noise levels of $q \in (0,1)$. Note that while normally distributed prediction errors often arise in regression-based predictive models are in many instances (cf. Schweigler et al. (2009), Sun et al. (2009)), time-series types of model would introduce dependencies across recent predictions; our noise model is intended to be a first-order approximation to understand the role noise plays.

4.1. Performance Under Noise

We compare the performance of the proposed $PA_w(s,l)$ policies against an online policy with a fixed threshold, $TH(L)$, under different levels of prediction noise. We start with such a benchmark policy as it closely mimics those used in practice (e.g., Allon et al. (2013)). The value of the threshold L is chosen such that the waiting time under the online policy is around 4 hours (cf. Batt and Terwiesch (2012)). A comparison to policies involving multiple thresholds will be discussed in Section 4.7.

The results are summarized in Table 1. We have chosen a diversion rate so that the average waiting time for the online threshold policy is around 4 hours. While some emergency departments have average waiting time less than 3 hours (Han et al. (2007), Mason et al. (2012)) they can also be quite large in other hospitals (Steele and Kiss (2008), Djokovic (2012)). With this in mind, we elected to consider a scenario in which the waiting is on the higher end, and prudent diversion policies are more necessary. As can be seen, the waiting time can be vastly improved with careful use of the available future information. Even when the information is quite noisy (e.g., $q = 0.9$), there are potential improvements which can be achieved via the proactive policy.

Noise	Online Wait (hrs)	Online Div.	Proac. Wait (hrs)	Proac. Div.	(s,l)	Impvmt.
$q = 0$	4.19	17.1%	3.55	17.1%	$(11,43)$	15.4%
$q = 0.1$	4.19	17.1%	3.78	17.0%	$(14,43)$	9.9%
$q = 0.5$	4.19	17.1%	3.85	17.0%	$(16,43)$	8.2%
$q = 0.9$	4.19	17.1%	3.88	17.0%	$(17,43)$	7.6%

Table 1 **Average waiting times under different levels of prediction noise, with $\tilde{\lambda} = 1.2$ and $w = 24$. In all cases, the online policy has a threshold of 42. The diversion rates are given as a percentage of the average arrival rate.**

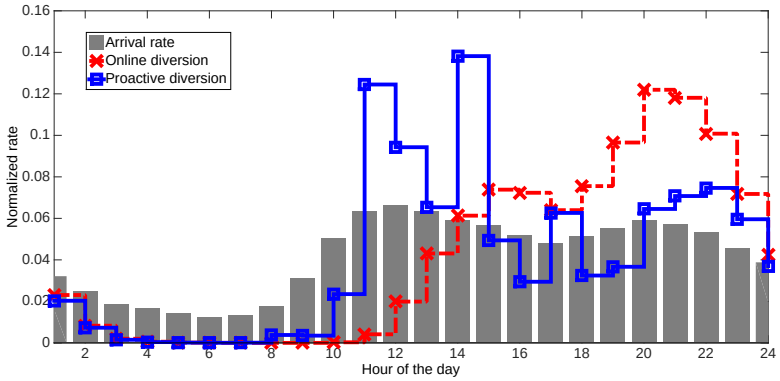


Figure 5 **Hourly rate of diversion averaged across all weekdays, with $q = 0.1$ and $(s, l) = (14, 43)$. All rates are normalized such that the area under curve is equal to 1. Notice the 'lag' of approximately 3 hours between the rise of the arrival rate (around 10 am) and when diversion rate of the online policy starts to increase (around 1 pm). The proactive policy, on the contrary, tends to 'foresee' the onset of heavier arrival intensities and increases its diversion effort shortly after 10 am.**

In addition to yielding a smaller average waiting time, the proactive policy also appears to improve the tail of the waiting time distribution. As shown in Table 2, the proactive policy yields smaller tail probabilities at all three levels of waiting time. Notably, the fraction of jobs that wait for more than 5 hours is reduced by 28%.

	Avg. wait (hrs)	P(Wait >5 hrs)	P(Wait >7 hrs)	P(Wait >9 hrs)
Online	4.19	39.1%	6.33%	0.25%
Proactive	3.78	28.2%	4.66%	0.21%

Table 2 **Comparison of tail probabilities of empirical waiting time distributions, with parameters identical to those used in Table 1 for $q = 0.1$.**

Figure 5 illustrates how the diversion rate pattern induced by the proactive policy can be significantly different from that of an online threshold policy. It also demonstrates why we refer to this class of policies as *proactive*. While the online policy waits for the congestion to get bad before reacting, the proactive policy is able to predict that the congestion is going to increase in the future and, so, proactively starts diverting patients.

We next look at the performance of the proactive policy as the amount of lookahead varies. As is shown in Table 3, while a longer lookahead window is generally better, our simulation results suggest that the marginal benefit tends to diminish as the window becomes large. This is consistent with our theoretical results. In practice, longer lookahead windows are likely to be associated

with noisier predictions. Thus, there is a fine tradeoff between decreasing delay by having more information versus risking diverting too many patients because there is too much noise.

Lookahead (hrs)	Proactive Wait (hrs)	(s,l)	Impvmt over Online
$w=8$	3.93	$(27,43)$	6.2%
$w=24$	3.78	$(14,43)$	9.9%
$w=48$	3.70	$(12,43)$	11.6%
$w=168$	3.62	$(11,43)$	13.7%

Table 3 **Average waiting times as the size of lookahead window varies, with $\tilde{\lambda}=1.2$ and $q=0.1$. The online policy has a threshold of 42, with a waiting time of 4.19 hours. and the diversion rate of the proactive policy is strictly smaller than that of the online policy for all cases.**

4.2. Sensitivity to Diversion Rate

Besides the prediction noise, another factor that will impact the system's waiting time performance is the rate of diversion, which corresponds to the parameter p in our theoretical model. Table 4 shows that the gain of the proactive policy over the online policy is fairly robust as the diversion rate changes. As the rate of diversion changes from 4% to 29% of the total arrival rate, the delay improvement of the proactive policy varies from 2% to 10%, with relatively larger improvement when the diversion rate is around 20%.

Arr. Rate	Online Div.	Online Wait (hrs)	Proa. Div.	Proa. Wait (hrs)	(s,l)	Impvmt.
$\tilde{\lambda}=0.98$	2.5%	3.19	2.5%	3.15	$(12,51)$	1.5%
$\tilde{\lambda}=1$	4.4%	3.05	4.4%	2.98	$(10,46)$	2.4%
$\tilde{\lambda}=1.1$	10.5%	4.13	10.4%	3.92	$(15,47)$	4.9%
$\tilde{\lambda}=1.2$	17.1%	4.19	17.0%	3.78	$(14,43)$	9.9%
$\tilde{\lambda}=1.3$	23.6%	4.16	23.6%	3.75	$(17,40)$	10.0%
$\tilde{\lambda}=1.4$	28.9%	4.12	28.8%	3.77	$(19,38)$	8.5%

Table 4 **Performance comparison among various levels of diversion rate, with $q=0.1$ and $w=24$. The arrival rates in the first column are expressed as a multiple of the system capacity. The online thresholds used are 50, 50, 50, 45, 46, 42, 39 and 37, for $\tilde{\lambda}$ equal to 0.8 through 1.4, respectively. Diversion rates are given as a percentage of the average arrival rate.**

These numerical results are consistent with our intuition. If the rate of diversion is too low relative to the system's traffic intensity, the degree of freedom of any feasible diversion policy may become too constrained to achieve large gains over the online policy. When $\tilde{\lambda}$ is close to or below 1, the proactive policy remains competitive against the online policy, though the delay improvement diminishes. For instance, when $\tilde{\lambda} = 0.98$, a proactive policy improves over the online policy by 1.5%. As the traffic intensity decreases further, it becomes difficult to identify substantial gains by using the proactive policy and we could not identify a scenario where the proactive policy achieved gains larger than 1%. This is likely because at lower loads, few patients need to be diverted in order to achieve low waiting times. For instance, when $\tilde{\lambda} = 0.9$ and 0.8, the average waiting time under an online policy with a threshold of 50 is only 2 and 0.92 hours with less than 0.7% of all arrivals diverted. Thus, there is little degree of freedom for further improvement. On the other extreme, if the diversion rate is so high so that a significant portion of the arrivals can be diverted, then even the online threshold policy is able to render a very small average queue length, thus making it difficult to achieve additional improvements through the use of predictive information.

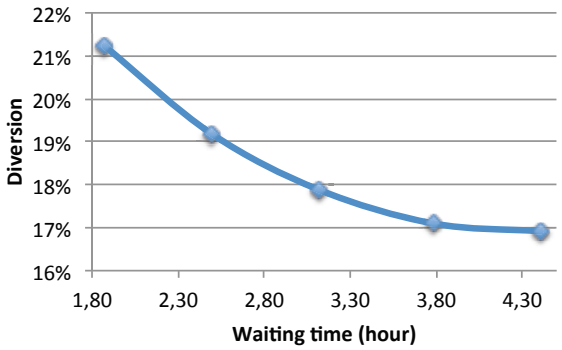


Figure 6 **Percentage of diverted arrivals and the corresponding waiting time, as the upper threshold, l increases, with $s = 14$, $\tilde{\lambda} = 1.2$, $q = 0.1$, and $w = 24$. From left to right, the data points correspond to $l = 22, 29, 36, 43$ and 50, respectively.**

Recall that the diversion rate, p, is a design parameter which can be tuned to achieve a particular performance. In particular, there is a tradeoff between more diversions and shorter waiting times. Thus, for a given arrival rate, the manager can vary the parameters of the proactive policy to achieve different diversion rate or waiting time targets. One way to do so is by changing the value of the upper and lower thresholds. Figure 6 shows the average waiting times along with their corresponding diversion rates as the upper threshold, l, varies.

4.3. Oscillating Diversion Status

We have thus far measured the diversion behavior of a policy mainly by the average diversion rate. In some scenarios, it may be advantageous not to constantly switch between diversion and

admission. For instance, this may be the case when diversion corresponds exclusively to diverting ambulances. In such a setting, it would be logistically challenging (and possibly infeasible) to convey to Emergency Medical Services drivers the frequently changing diversion status of the hospital.

To gain some insight into the frequency of alternating between diverting and not diverting patients, we consider the duration spent on- versus off-diversion. More precisely, we denote by S_{on} the average time span between the first diversion of a sequence until the next time a patient is admitted to the system. Thus, any patient who arrived to the system during this time span would have been diverted. Likewise, we denote by S_{off} as the average time span between the first admission of a sequence until the next time a patient is diverted from the system. Note that a sequence can consist of a single patient if there are frequent oscillations between on- and off-diversion. We would like both values to be large to indicate infrequent oscillations. With identical parameters as in Table 1 for $q = 0.1$, simulations show that both S_{on} and S_{off} are substantially greater under the proactive policy (0.16 and 2.41 hours, respectively), compared to the online policy (0.07 and 1.94 hours, respectively), indicating that it makes less frequent switches. Additionally, we find that these times spent on diversion are consistent with what is seen in the ED literature (e.g., Kahn et al. (2014)).

4.4. Heterogeneous Patient Types

The proactive policy can also be adapted to a system with service priorities, and where diversions can be restricted only to a subset of priority classes. Similar to Allon et al. (2013) and Deo and Gurvich (2011), we consider a two-class priority system, where every arrival belongs to a high-priority class with probability p_h, and is of low priority, otherwise. The high priority class may correspond to ambulance arrivals while the low priority class would correspond to walk-ins. An available server will always first choose a high-priority job if there is any in the system. We further consider two diversion scenarios, where the manager is allowed to divert only arrivals from one of the two priority classes. The case where only low-priority jobs are diverted can be thought of as referrals of low-acuity patients to PCP or urgent care, and the other case corresponds to ambulance diversions of high-acuity patients. The proactive policy for the priority system, further elaborated on in Appendix B.2, is essentially identical to the original except that the computation of the base-line queue length process now takes into account the priority service rule.

The results are summarized in Table 5. When low-priority jobs are diverted, the proactive policy achieves an average delay improvement of 9.4%, or 24 minutes, while essentially maintaining the same high-priority waiting time, which is increased by 0.7%, or 2.1 seconds. When only high-priority jobs are diverted, the performance improvement is small when the fraction of such jobs, p_h, is small.

Nevertheless, when $p_h = 0.3$ (e.g., the fraction of ambulance arrivals in a Hong Kong ED according to Xu et al. (2013)), the proactive policy is able to improve the waiting time of the high-priority class by 15%, while minutely increasing the waiting time of the low-priority class (+0.2%). As the fraction of high-priority jobs becomes more significant, e.g., $p_h = 0.7$, the proactive policy is able to decrease waiting time in both classes, and drastically reducing waiting in the high-priority class, by as much as 26.9%. Similar to Table 2, we also observe, in all three cases, improved waiting time tail probabilities under the proactive policy in the priority class where diversions are allowed. We expect similar results to hold if the number of classes increases.

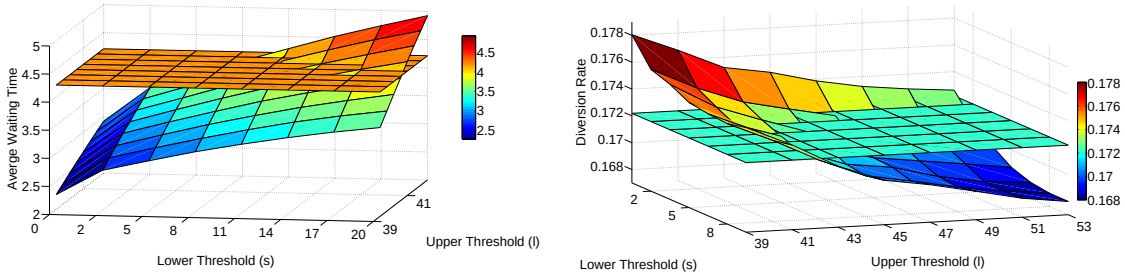
Div. scenario	p_h	High-priority wait (hrs)	Low-priority wait	Average wait
Low-priority	0.1	0.08 (+0.74%)	4.38 (−9.46%)	3.85 (−9.43%)
High-priority	0.3	0.12 (−15.2%)	4.53 (+0.23%)	3.20 (+0.06%)
High-priority	0.7	0.30 (−26.9%)	5.18 (−2.23%)	2.13 (−5.25%)

Table 5 **Performance of proactive policy in systems with priority, with $\tilde{\lambda} = 1.2$, $q = 0.1$, and $w = 24$. The values in the parentheses are the percentage change as compared to an online threshold policy. For all three scenarios, the diversions rates of the proactive policy are strictly smaller than those of the online policy. The thresholds for the online policy are 42, 24, and 24, and for the proactive policy $(s, l) = (14, 43), (1, 25)$ and $(1, 25)$, respectively.**

4.5. Choosing (s, l)

In our simulations the lower and upper thresholds for the proactive policy, (s, l), were chosen according to the following simple procedure: we fix the upper threshold, l, to be 1 plus the online threshold, and then find the smallest lower threshold, s, under which the diversion rate of the proactive policy is no greater than that of the online policy. Note that, in the setting of Theorem 1, with a noise-less lookahead window of infinite length, it suffices to set l to be *equal* to the online threshold, and s being equal to zero. However, it appears that the presence of time-varying arrival rates, noise, and the finiteness of the lookahead window, requires that our upper threshold to be strictly greater than the online threshold, and the lower threshold strictly greater than zero, in order to avoid excessive diversion.

Figures 7(a) and 7(b) illustrate the behavior of the delay and diversion rate of the proactive policy as a function of the lower and upper thresholds, obtained via simulations, with $q = 0$. It appears that delay decreases as both thresholds decrease, and the opposite is true for the diversion



(a)Average waiting time of the proactive policy as a function of the lower and upper thresholds.

(b)Rate of diversion, as a percentage of total arrival rate, for the proactive policy, as a function of the lower and upper thresholds.

Figure 7 Dependency on the lower and upper thresholds (s,l) of the Proactive Policy's performance, with $q=0$ and $\tilde{\lambda}=1.2$. The flat surface corresponds to the performance of the online policy.

rate. This is consistent with our intuition of the roles of the thresholds as discussed in previous sections. Through exhaustive search, we have found that our relatively low complexity procedure for choosing (s,l) can be sub-optimal, and it may be possible to further improve our result via an exhaustive search over a larger range of s and l. Indeed, for the case of $q=0$, we can find another pair with a higher upper threshold: $(s,l)=(5,47)$, which yields a delay improvement of 18.6% over the online setting. This is an additional 3% improvement over the choice of (s,l) in Table 1, while it is relatively small compared to the 15.4% delay improvement already achieved. Similar phenomena appear to hold for other values of q as well, in which one can typically achieve a delay improvement of up to 3% by performing an exhaustive search. Unfortunately, the process of exhaustive search currently relies on time-consuming simulations and is not very scalable. The simple procedure of choosing (s,l) mentioned above speeds up the process by constraining the search onto one dimension, and appears to produce near optimal delay performances.

4.6. Impact of Left Without Being Seen (LWBS)

It has also been reported in the literature that extended emergency room delays can lead to patients leaving the hospital without being seen by a physician (LWBS) (Green et al. (2006), Batt and Terwiesch (2015), Bolandifar et al. (2013)). To model this phenomenon, we assume that, in each time slot, each job in queue has a constant probability of abandonment, γ, and the value of γ is chosen so that the rate of LWBS for the online policy is around 10% of the average arrival rate (cf. Green et al. (2006)). We verify via simulations that the benefits of the proactive policy continue to hold in a setting where admitted patients may abandon while waiting in queue. In particular, in the setting illustrated in Table 6, the proactive policy is able to improve delay by 7.7%, while ensuring that the total rate of diversion and LWBS does not exceed that of the online policy.

	Waiting (hrs)	Diversion	LWBS	Diversion + LWBS
Online	2.87	7.2 %	11.5 %	18.7 %
Proactive	2.65 (-7.7 % from online)	8.0 %	10.6 %	18.6 %

Table 6 **Comparison of online and proactive policies with $\tilde{\lambda} = 1.2$, $q = 0.1$, and $w = 24$, under patient abandonment. The proactive policy achieves a 7.7% delay improvement, with a total rate of diversion and LWBS that is no greater than the online policy. The threshold for the online policy is 42; for the proactive policy they are 14 and 52. Diversion rates are given as a percentage of the average arrival rate.**

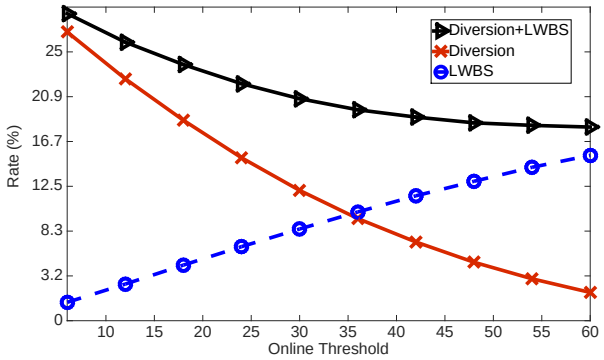


Figure 8 **Rate of diversion and LWBS as a function of the online threshold.**

Remark 1 *Even though the proactive policy is able to serve more patients (i.e. lower total rate of diversion and LWBS) under this* specific *choice of online threshold, one may wonder whether a* smaller *online threshold could lead to smaller combined rate of diversion and LWBS, hence beating the proactive policy. Figure 8 shows that this is not possible, as the combined rate of diversion and LWBS for the online policy is* monotonically increasing *as the online threshold decreases, even as the diversion and LWBS rates themselves obey opposite patterns. In an ideal world, one may want to find a policy that incurs* both *less diversion* and *less LWBS when compared to the online policy, while delivering a smaller delay. However, we have not been able to identify a set of parameters under which this stronger notion of improvement can be achieved with the proactive policy.*

4.7. Comparison to Multi-Threshold Policies

Multi-threshold policies are generalizations of the online threshold policies, where the value of the threshold may vary as a function of time. While such policies are 'online' in the sense that they do not use predictions of future arrivals, but rather leverage known variation in arrival rates over the course of a day, this approach has several drawbacks. Firstly, a time-varying threshold is not the approach which is typically used in most hospitals. As discussed in Allon et al. (2013) and references therein, hospitals will typically go on ambulance diversion when the number of

patients waiting (and/or boarding) exceeds a single threshold. Also, it is unclear how to choose the thresholds using the arrival rate information in a principled manner, and a brute-forced search can quickly become computationally intractable as the set of all possible choices grows exponentially in the number of potential thresholds. Finally, unlike the proactive policy studied in this paper, a time-varying threshold based on arrival rates does not lend itself easily to incorporating additional future information, should it become available. For these reasons, we have been focusing on a fixed threshold in our simulations.

Nevertheless, it remains of interest to know whether adopting multiple thresholds could vastly improve performance over the single-threshold policy, which is no longer optimal under time-varying arrival rates, and how the improvement compares to that achieved by the proactive policies. While a comprehensive investigation of multi-threshold policies for our admission control problem is beyond the scope of this paper, as a first step, we examine a family of two-threshold policies, denoted by $MTH(L_a, L_b)$ which applies a threshold L_a during the interval of 11:00 am through 8:59 pm, and a second threshold L_b for the interval of 9:00 pm through 10:59 am. The allocation of the intervals is based on the observation that the hourly arrival rates from 11 am till 8 pm tend to be substantially greater than those during the remainder of the day. We will use the online single-threshold policy $TH(L)$, with $L = 42$ and $\tilde{\lambda}$=1.2 (Table 1) as a performance benchmark.

We exhaustively search over simulations of all pairs of $L_a, L_b \in \{27, \ldots, 57\}$ (a 15-by-15 square grid centered around the benchmark single threshold of 42). Among all feasible pairs of (L_a, L_b), whose diversion rate is no greater than that of the $TH(42)$ policy, the pair $(L_a, L_b) = (44, 36)$ achieves the maximum, 2.78%, reduction in average queue length compared to that of $TH(42)$. This is substantially smaller compared to the $7.6\% \sim 15.4\%$ queue length reductions achieved by the proactive policies under the same set of parameters and constraints (Table 1).

The aforementioned example does not rule out the possibility of further improvement by enlarging the number of thresholds and their temporal arrangements. Nevertheless, the relatively moderate delay improvement from the $MTH(L_a, L_b)$ policies serves as further evidence that the proactive policy may be capable of achieving superior performance even when compared to policies that are tailored to time-varying arrival rates.

5. Conclusions and Discussions

This paper provides an important first step at understanding how to reduce waiting times by making admission decisions which utilize predictions of future arrivals. There has been substantial attention in the medical and health service literature paid towards developing predictive models;

yet, thus far, there has been very limited work done to understand how to utilize such tools to improve patient flow. This work provides insights towards that goal.

We developed a framework for understanding when and how *predictive information* can be used to improve waiting times in service settings. Our primary motivation is the Emergency Department where waiting can degrade patient outcomes and 'diversion' of patients to other care facilities may be possible. We proposed a family of proactive diversion policies to leverage (noisy) predictions of future arrival and service patterns. We demonstrated that, given sufficient future information, the proactive policies were able to achieve superior waiting time performance over the optimal online policy at any traffic intensity, and provided bounds on the performance of the policy for the regime where the future information is more limited. We investigated the impact of prediction noise, and identified a noise tolerance range, within which the proactive policies will provide essentially the same performance guarantees despite having imperfect predictions, yet beyond which the policy will make excessive diversions due to the overwhelming prediction noise. A key observation from our analysis is that, instead of waiting for congestion to build up, a manager should be *proactive* and divert more aggressively towards the beginning of a 'bursty episode', where there will be significantly more arrivals than service availabilities.

The effectiveness of the proposed proactive diversion policies were examined in the ED context through numerical simulations. The simulation results show that the proactive policies, despite their structural simplicity, consistently outperform online threshold policies across a variety of settings, with up to 15% reduction in average waiting time. The proactive policies are also robust, in the sense that the performance improvement is maintained as we vary the noise level, rate of diversion, and rate of patient abandonment.

Our proactive policy is likely not the only way to leverage predictive information, and it will be interesting to study other policies and understand their relative strengths under different traffic intensities or levels of prediction noise. For instance, as was mentioned in Section 4, an online threshold policy with a time-varying threshold that depends on the mean arrival rates throughout the day could be a viable alternative when the quality of future information is poor and the rate of change of the arrival rates is relatively small. Also, while our proactive policy yields superior delay improvement over an optimal online policy when there is sufficient future information, it remains an open problem whether it is possible to find a proactive policy that provably outperforms the online policies even when the length of the lookahead is very small. Additionally, it would be interesting to develop a rigorous understanding of the regime of diversion rates in which the proactive policy will perform the best. We leave this exploration for future work.

Our current queueing model assumes that diversions can be made at any time of the day, while in practice there can be off-periods during which no diversion can be made. In a related vein, Deo and Gurvich (2011) demonstrated that in some instances when one hospital goes on diversion, it can cause all other hospitals to go on diversion, effectively making it so no one can be on diversion. Incorporation of these additional diversion constraints into the analytically framework can be an interesting topic for future research.

Finally, the prediction-guided diversion model studied in this paper is quite general and may also be applicable to other areas of service, and more specifically, health-care operations. There has been a lot of attention on the development of predictive models in health-care. For instance, Armony et al. (2015) shows that the discharge times from an internal ward can be highly predictable, and Bayati et al. (2014) has developed a host of predictive models for hospital readmission. It will be interesting to further understand how such models can be leveraged in improving operational efficiency and performance.

References

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

American Hospital Association. 2010. Percent of hospitals reporting emergency department capacity issues by type of hospital, 2007. *Rapid Response Survey, Telling the Hospital Story* .

Armony, N., S. Israelit, S. Mandelbaum, Y. Marmor, Y. Tseytlin, G. Yom-Tov. 2015. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems, to appear* .

Batt, R.J., C. Terwiesch. 2012. Doctors under load: An empirical study of state-dependent service times in emergency care. *Working Paper, The Wharton School* .

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

Bayati, M., M. Braverman, M. Gillam, K. Mack, G. Ruiz, M. Smith, E. Horvitz. 2014. Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PLoS ONE* **9**(10) e109264.

Bolandifar, E., N. DeHoratius, T. Olsen, J. Wiler. 2013. An econometric analysis of patients who leave the ed without being seen by a physician. *Working Paper, The Chinese University of Hong Kong* .

Burt, C.W., S.M. Schappert. 2004. Ambulatory care visits to physician offices, hospital outpatient departments, and emergency departments: United States, 1999-2000. *Vital Health Stat.* **13**(157) 1–70.

Carr, M., B. Hajek. 1993. Scheduling with asynchronous service opportunities with applications to multiple satellite systems. *IEEE Trans. Automatic Control* **38**(12) 1820–1833.

Chase, V.J., A.E. Cohn, T.A. Peterson, M.S. Lavieri. 2012. Predicting emergency department volume using forecasting methods to create a "surge response" for noncrisis events. *Academic Emergency Medicine* **19**(5) 569–576.

Committee on the Future of Emergency Care in the United States. 2007. *Emergency medical services at the crossroads*. Washington, DC: The National Academies Press.

Deo, S., I. Gurvich. 2011. Centralized vs. Decentralized Ambulance Diversion: A Network Perspective. *Management Science* **57**(7) 1300–1319.

Djokovic, M. 2012. Increased emergency department boarding times. *UMass Amherst Doctor of Nursing Practice (DNP) Capstone Projects* .

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

Gans, N., H. Shen, Y.P. Zhou, N. Korolev, A. McCord, H. Ristock. 2015. Parametric forecasting and stochastic programming models for call-center workforce scheduling. *Manufacturing and Service Operations Management, to appear* .

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

Gross, Donald, John F Shortle, James M Thompson, Carl M Harris. 2013. *Fundamentals of queueing theory*. John Wiley & Sons.

Guarisco, J., . A. Samuelson. 2011. Rx for the ER. *OR/MS Today* **38**(5) 32–35.

Han, J. H., C. Zhou, D. J. France, S. Zhong, I. Jones, A. B. Storrow, D. Aronsky. 2007. The effect of emergency department expansion on emergency department overcrowding. *Academic Emergency Medicine* **14**(4) 338–343.

Helm, J. E., S. AhmadBeygi, M. P. Van Oyen. 2011. Design and Analysis of Hospital Admission Control for Operational Effectiveness. *Production and Operations Management* **20**(2) 359–374.

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

Japsen, B. 2003. Hospital capacity debate heats up: Aging population means sharp rise in need, study says. *Chicago Tribune. July 17* .

Jones, S.A., M.P. Joy, J. Pearson. 2002. Forecasting demand of emergency care. *Health Care Management Science* **5**(4) 297–305.

Jones, S.S., R.S. Evans, T.L. Allen, A. Thomas, P.J. Haug, S.J. Welch, G.L. Snow. 2009. A multivariate time series approach to modeling and forecasting demand in the emergency department. *Journal of Biomedical Informatics* **42**(1) 123–139.

Kahn, C. A., S. J. Stratton, C. L. Anderson. 2014. Characteristics of Hospitals Diverting Ambulances in a California EMS System. *Prehospital and Disaster Medicine* **29**(1) 27–31.

Khare, R., E. Powell, G. Reinhardt, M. Lucenti. 2009. Adding more beds to the emergency department or reducing admitted patient boarding times: which has a more significant influence on emergency department congestion? *Annals of emergency medicine* **53**(5) 575–585.

Kolker, A. 2008. Process modeling of emergency department patient flow: Effect of patient length of stay on ED diversion. *Journal of Medical Systems* **32**(5) 389–401.

Mason, S., E. J. Webe, J. Coster, J. Freeman, T. Locker. 2012. Time patients spend in the emergency department: England's 4-hour rulea case of hitting the target but missing the point? *Annals of Emergency Medicine* **59**(5) 341–349.

McCarthy, M.L., S.L. Zeger, R. Ding, D. Aronsky, N.R. Hoot, G.D. Kelen. 2008. The challenge of predicting demand for emergency department services. *Academic Emergency Medicine* **15**(4) 337–346.

McCusker, J., J. Verdon. 2006. Do geriatric interventions reduce emergency department visits? a systematic review. *Journal of Gerontology* **61**(1) 53–62.

Nawijin, N.W. 1990. Look-ahead policies for admission to a single server loss system. *Operations Research* **38**(5) 854–862.

Peck, Jordan S, Stephan A Gaehde, Deborah J Nightingale, David Y Gelman, David S Huckins, Mark F Lemons, Eric W Dickson, James C Benneyan. 2013. Generalizability of a simple approach for predicting hospital admission from an emergency department. *Academic Emergency Medicine* **20**(11) 1156–1163.

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

Riegel, B., B. Carlson, Z. Kopp, B. LePetri, D. Glaser, A. Unger. 2002. Effect of a standardized nurse case-management telephone intervention on resource use in patients with chronic heart failure. *Archives of Internal Medicine* **162**(6) 705–712.

Rotstein, Z., R. Wilf-Miron, B. Lavi, A. Shahar, U. Gabbay, S. Noy. 1997. The dynamics of patient visits to a public hospital ed: a statistical model. *Am J Emerg Med* **15**(6) 596–9.

Saghafian, S., W. Hopp, M. Van Oyen, J. Desmond, S.L. Kronick. 2012. Patient Streaming as a Mechanism for Improving Responsiveness in Emergency Departments. *Operations Research* **60**(5) 1080–1097.

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

Schweigler, L.M., J.S. Desmond, M.L. McCarthy, K.J. Bukowski, E.L. Ionides EL, J.G. Younger. 2009. Forecasting models of emergency department crowding. *Academic Emergency Medicine* **16**(4) 301–308.

SEE-Center, Technion. 2009. Seestat database. URL `http://seeserver.iem.technion.ac.il/see-terminal/`.

Spencer, J., M. Sudan, K. Xu. 2014. Queuing with future information. *The Annals of Applied Probability* **24**(5) 2091–2142.

Steele, R., A. Kiss. 2008. Emdoc (emergency department overcrowding) internet-based safety net research. *The Journal of Emergency Medicine* **35**(1) 101–107.

Stidham, S.Jr. 1985. Optimal control of admission to a queueing system. *IEEE Trans. Automatic Control* **30**(8) 705–713.

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

Sun, Y., B. H. Heng, Y. T. Seow, E. Seow. 2009. Forecasting daily attendances at an emergency department to aid resource planning. *BMC Emergency Medicine* **9**(1).

Tandberg, D., C. Qualls. 1994. Time series forecasts of emergency department patient volume, length of stay, and acuity. *Annals of emergency medicine* **23**(2) 299–306.

Wargon, M., B. Guidet, T.D. Hoang, G. Hejblum. 2009. A systematic review of models for forecasting the number of emergency department visits. *Emergency Medicine Journal* **26**(6) 395–399.

Xu, K. 2015. Necessity of future information in admission control. *to appear in Operations Research* .

Xu, M., T.C. Wong, S.Y. Wong, K.S. Chin, K.L. Tsui, R.Y. Hsia. 2013. Delays in service for non-emergent patients due to arrival of emergent patients in the emergency department: A case study in hong kong. *The Journal of Emergency Medicine* **45**(2) 271–280.

Appendix

A. Miscellaneous Proofs

A.1. Proof of Theorem 1

PROOF OF THEOREM 1: The main idea of the proof is to characterize the queue length process induced by a $PA_\infty(0,l)$ policy as a truncated birth-death process, from which both the diversion rate and the steady-state expected queue length can be derived. This characterization in turn hinges on a technical result (Lemma 1), which shows that the $PA_w(0,l)$ policy can be 'sequentialized' into two separate steps, without changing the resulting queue length. We refer to this policy as the $\widehat{PA}_w(0,l)$ policy. Compared to the original $PA_w(0,l)$ policy, the two-step policy, given in Definition 3, first diverts all w-blocking arrivals, before 're-running' the system and then diverting among the remaining jobs those that arrive when $Q(t)=l$. The analysis of this equivalent two-step policy turns out to be easier than the original version, for it allows one to disentangle the effect of the diversions of w-blocking arrivals from that of the diversions made by thresholding.

Definition 3 *The policy $\widehat{PA}_w(0,l)$ consists of the following two steps.*

1. *Step 1: Apply the policy $PA_w(0,\infty)$ to the baseline queue length process; that is, divert all w-blocking arrivals. Define A_{PA_∞} to be the counting process that consists of all arrivals remaining, i.e., those that are* not *diverted under the $PA_w(0,\infty)$ policy.*
2. *Step 2: Apply an online threshold policy, with threshold l, to a system with arrival process A_{PA_∞}, service token process S, and an initially empty queue.*

The next lemma shows that, for every realization of the arrival and service token processes, the $\widehat{PA}_w(0,l)$ policy is equivalent to the original $PA_w(0,l)$ policy in that they produce the same set of diversions, and consequently, the same resulting queue length process. The proof of the result is given in Appendix A.4.

Lemma 1 *Fix $l\in\mathbb{Z}_+$ and $w\in\mathbb{R}_+\cup\{\infty\}$. Denote by D and $\widehat{D}\subset A$ the sets of diversions made by $PA_w(0,l)$ and $\widehat{PA}_w(0,l)$, respectively. Then, $D=\widehat{D}$ almost surely.*

In light of Lemma 1, we can focus on the $\widehat{PA}_\infty(0,l)$ policy for the remainder of the proof of Theorem 1. The main benefit of analyzing this two-step policy is that we can obtain a full characterization of the queue length process Q_{PA_∞}, induced by the first step, using the following result from Spencer et al. (2014).

Proposition 1 *(Adapted from Proposition 1, Spencer et al. (2014)) For all $\lambda\in(\max\{p,1-p\},1)$, Q_{PA_∞} is a positive recurrent birth-death process, whose sample paths are distributed according to that of the total number of jobs in an initially empty $M/M/1$ queue with arrival rate $1-p$ and service rate λ.*

We are now ready to show the steady-state distribution of the queue length process under $PA_\infty(0,l)$, in Eq. (5). Using Proposition 1, the second step of $\widehat{PA}_\infty(0,l)$ effectively truncates the birth-death process associated with Q_{PA_∞} at state l, leading to a Markov chain whose transition rates are illustrated in Figure

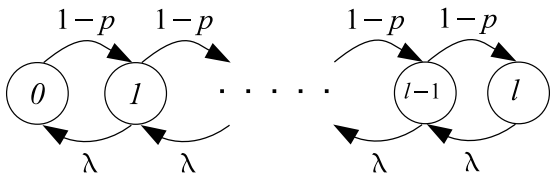


Figure 9 **The transition rates of the continuous-time Markov chain that corresponds to the queue length process after applying the $PA_\infty(0,l)$ policy. The transition rates are identical to those induced by applying an l-threshold online policy to a queue with arrival rates $1-p$ and service rate λ. Note that the transition rates of the Markov chain that corresponds to the online threshold policy $TH(l)$ can be obtained from this diagram by changing all $1-p$ to λ, and all λ to $1-p$, i.e. the transition rates are flipped.**

9. The expressions for the steady-state distribution of this chain follow from the standard techniques, which involve solving a set of balance equations specified by the transition rates. This proves Eq. (5).

We next show the expression of the optimal threshold l for the $PA_\infty(0,l)$ policy, given in Eq. (6). From the definition of $\widehat{PA}_\infty(0,l)$, it is not difficult to see that the expected queue length decreases as the threshold l decreases, as a result of the thresholding in Step 2. Therefore, to find the optimal threshold that leads to the minimum expected queue length, it suffices to find a smallest l, under which the total rate of diversion from *both* steps of $\widehat{PA}_\infty(0,l)$ does not exceed p.

To this end, we compute the diversion rates induced by the two steps of $\widehat{PA}_\infty(0,l)$ separately, which we denote by d_1 and d_2, respectively. For the first step, it is not difficult to show that the diversion of all ∞-blocking arrivals amounts to diverting all arrivals that would have not been processed by the server in finite time (cf. Lemma 2, Spencer et al. (2014)). This leads to a diversion rate of

$$d_1 = \lambda - (1-p), \tag{16}$$

which is equal to the discrepancy between the arrival and service rates. For the second step of $\widehat{PA}_\infty(0,l)$, note that an arrival is diverted if any only if the process Q_{PA_∞} is in state l. Furthermore, the rate of birth in Q_{PA_∞} is equal to $1-p$, by Proposition 1. We thus have that the diversion rate induced by the second step is given by

$$d_2 = \pi_l(1-p), \tag{17}$$

where π_l is the steady-state probability of the queue being in state l under $\widehat{PA}_\infty(0,l)$, with $\pi_l = \frac{1-\beta}{1-\beta^{l+1}}\beta^l$, with $\beta = \frac{1-p}{\lambda}$ (Eq. (5)).

Combining the above diversion rates for the two steps in $\widehat{PA}_\infty(0,l)$, Eqs. (16) and (17), it suffices to choose the smallest l for which the following holds:

$$d_1 + d_2 = [\lambda - (1-p)] + \frac{1-\beta}{1-\beta^{l+1}}\beta^l(1-p) \le p, \tag{18}$$

which yields the requirement

$$l \ge \log_{\frac{\lambda}{1-p}} \frac{p}{1-\lambda} - 1 = L(p,\lambda). \tag{19}$$

Hence, $l^* = L(p,\lambda)$. This proves Eq. (6).

For Eq. (7), the expected queue length under $\widehat{PA}_\infty(0,l^*)$ can be readily computed from the steady-state probabilities in Eq. (5) and the value of l^*. Using Eq. (5), we have that

$$\mathbb{E}\left(Q_{PA^*}\right) = \sum_{i=1}^{L(p,\lambda)} i\pi_i = \frac{1}{1-\beta^{-(L(p,\lambda)+1)}} L(p,\lambda) - \frac{\beta(\beta^{L(p,\lambda)}-1)}{(\beta-1)(\beta^{L(p,\lambda)+1}-1)}. \tag{20}$$

Combining this with the fact that $\beta^{L(p,\lambda)+1} = \left(\frac{1-p}{\lambda}\right)^{\left(\log_{\frac{\lambda}{1-p}} \frac{p}{1-\lambda}\right)} = \frac{1-\lambda}{p}$ yields Eq. (7).

We now show Eq. (8). Recall that the optimal expected queue length in the online setting can be achieved by a threshold policy, $TH(L)$, where $L = L(p,\lambda) = \log_{\frac{\lambda}{1-p}} \frac{p}{1-\lambda}$ (Eq. (1)). Therefore, the equality in Eq. (8) follows by combining Eq. (7) and the expression for the expected queue length under the $TH(L(p,\lambda))$ policy in Eq. (2).

Finally, to show that $\mathbb{E}(Q_{PA^*})$ is strictly smaller than $\mathbb{E}(Q_{ON})$ whenever $\lambda \in (\max\{p,1-p\},1)$, it is helpful to go back to the steady-state queue length distributions induced by the two diversion policies. Recall from Eq. (1) that the optimal expected queue length in the online setting is achieved by the threshold policy $TH(L)$, with threshold $L = L(p,\lambda) = l^*$. Denote by ψ_i the probability of $Q = i$ under the online policy $TH(l^*)$. By definition, the online policy $TH(l^*)$ induces a birth-death process truncated at state l^*, with rates of birth and death given by λ and $1-p$, respectively. Again, via solving the associated balancing equations, we have

$$\psi_i = \begin{cases} \frac{\beta^l(1-\beta)}{1-\beta^{l+1}}\beta^{-i}, & 0 \le i \le l^*, \\ 0, & \text{otherwise}, \end{cases} \tag{21}$$

where $\beta = \frac{1-p}{\lambda}$. Note that the queue length distribution for the $\widehat{PA}_\infty(0,l^*)$ policy, $\{\pi_i\}$, is a mirror image of that for the online policy, $\{\psi_i\}$, reflected across $l^*/2$.

Combining Eq. (21) with the expressions for π_i (Eq. (5)) and the fact that $\beta < 1$ whenever $\lambda > \max\{p,1-p\}$, we conclude that ψ_i and π_i are monotonically *increasing* and *decreasing* in i, respectively. This implies that

$$\mathbb{E}(Q_{ON}) = \sum_{i=1}^{l^*} i\psi_i > \frac{1}{2}l^* > \sum_{i=1}^{l^*} i\pi_i = \mathbb{E}(Q_{PA^*}), \tag{22}$$

where the two inequalities follow from the monotonicities of ψ_i and π_i, respectively. This proves the inequality in Eq. (8).

Finally, Eqs. 9 and 10 follow directly from Little's Law by recognizing the arrival rate of admitted jobs is $\lambda - p$. This completes the proof of Theorem 1. □

A.2. Proof of Theorem 2

Proof of Theorem 2: Because of the presence of no-shows, we shall distinguish between the set of *attempted diversions* in A', made by a proactive policy, and the set of *realized diversions* in A, after the no-shows have been realized. The diversion rate constraint applies only to the realized diversions.

We first consider the case where $l < \infty$. We will again analyze the two-step policy $\widehat{PA}_w(0,l)$, defined in Definition 3. Because the realizations of the no-shows do not depend on diversion actions, one can show that the $\widehat{PA}_w(0,l)$ policy produces the same queue length sample path as the original $PA_w(0,l)$ policy, just like in the noiseless setting of Theorem 1. This is stated in the following lemma, whose proof is similar to that of Lemma 1 and is omitted.

Lemma 2 *Fix $l \in \mathbb{Z}_+$ and $w \in \mathbb{R}_+ \cup \{\infty\}$. Denote by D and $\widehat{D} \subset A$ the sets of realized diversions induced by $PA_w(0,l)$ and $\widehat{PA}_w(0,l)$, respectively, under the no-show noise model. Then, for any $\epsilon > 0$, $D = \widehat{D}$ almost surely.*

By Lemma 2, we will focus on the $\widehat{PA}_w(0,l)$ for the remainder of the proof. We now compute the resulting diversion rate of the $\widehat{PA}_w(0,l)$ policy. Denote by d_1 and d_2 the rates of *realized* diversions induced by the first and second step of the $\widehat{PA}_\infty(0,l)$ policy, respectively. The first step of $\widehat{PA}_\infty(0,l)$ corresponds to applying a $PA_\infty(0,\infty)$ policy to a system with arrival process A' and service token process S. Using the same argument as the one proceeding Eq. (16), we have that the rate of *attempted* diversions among A' is equal to $\left[\frac{\lambda}{1-\epsilon} - (1-p)\right]$, i.e., the discrepancy between the rates of A' and S. Because each of the attempted diversions has a probability of ϵ of being a no-show, independently of all other diversions, we have that the rate of realized diversions

$$d_1 = \left[\frac{\lambda}{1-\epsilon} - (1-p)\right](1-\epsilon) = \lambda - (1-\epsilon)(1-p). \tag{23}$$

We next characterize the rate of realized diversions, d_2, induced by the second step of $\widehat{PA}_\infty(0,l)$. To facilitate our discussion, we will use the notation $Q(A,S)$ to denote the queue length process for a system that is initially empty at time $t=0$, with arrival process A and service token process S. Denote by D'_1 the set of attempted diversions made by the first step of $\widehat{PA}_w(0,l)$. Let A'_{PA_∞} be the set of *potential* arrivals after D'_1 had been removed from A':

$$A'_{PA_\infty} = A' \backslash D'_1, \tag{24}$$

and let A_{PA_∞} be the set of *realized* arrivals in A'_{PA_∞}. By applying Proposition 1, we conclude that $Q(A'_{PA_\infty}, S)$ admits the same distribution as that of an $M/M/1$ queue with arrival rate $1-p$ and service rate $\lambda/(1-\epsilon)$. Because each potential arrival in A'_{PA_∞} has probability ϵ of being a no-show, independently of all other potential arrivals, we have that the process $Q(A_{PA_\infty}, S)$, induced by the *realized* arrival process A_{PA_∞}, corresponds to the that of an $M/M/1$ queue with arrival rate $(1-p)(1-\epsilon)$ and service rate $\lambda/(1-\epsilon)$. Finally, thresholding the process $Q(A_{PA_\infty}, S)$ at length l yields the realized rate of diversion

$$d_2 = (1-p)(1-\epsilon)\pi_l = (1-\epsilon)(1-p)\frac{1-\beta}{1-\beta^{l+1}}\beta^l, \tag{25}$$

where $\pi_l = \frac{1-\beta}{1-\beta^{l+1}}\beta^l$ is the steady-state probability of being in l for the thresholded queue length process (cf. Eq. (5)), and

$$\beta = \frac{(1-p)(1-\epsilon)}{\lambda/(1-\epsilon)} = (1-\epsilon)^2\frac{\max\{p,1-p\}}{\lambda}. \tag{26}$$

Combining Eqs. (23) and (25), we have that the total rate of realized diversions induced by $\widehat{PA}_\infty(0,l)$ is given by

$$\begin{aligned} d =& d_1 + d_2 = \lambda - (1-\epsilon)(1-p) + (1-\epsilon)(1-p)\frac{1-\beta}{1-\beta^{l+1}}\beta^l \\ =& \lambda - (1-\epsilon)(1-p)\left(1 - \frac{1-\beta}{1-\beta^{l+1}}\beta^l\right) \end{aligned}$$

$$=\lambda-(1-\epsilon)(1-p)\frac{1-\beta^l}{1-\beta^{l+1}}, \tag{27}$$

which, combined with the diversion rate constraint of $d \leq p$, yields Eq. (11).

Now suppose that $l=\infty$. We note that because there does not exist a second step of thresholding for the $\widehat{PA}_w(0,\infty)$ policy, the resulting total realized diversion rate is given by

$$d = d_1 = \lambda-(1-\epsilon)(1-p), \tag{28}$$

and Eq. (11) follows from the constraint that $d \leq p$. Finally, when $l=\infty$, the expression for the expected steady-state queue length in Eq. (12) corresponds to the well-known steady-state expected value of a birth-death process on $\mathbb{Z}_+$, where the birth rate is $(1-p)(1-\epsilon)$ and death rate is $\lambda/(1-\epsilon)$. Similarly, when $l<\infty$, Eq. (12) corresponds to the expected value of the same birth-death process truncated at level l. This completes the proof of Theorem 2. □

A.3. Proof of Theorem 3

Proof of Theorem 3: Denote by D_1 the set of all w-blocking arrivals in A. We will follow steps similar to those in the proof of Theorem 1, with the additional task of keeping track of the extra diversion rate caused by the finite length of the lookahead window. In particular, we will analyze the two-step policy, $\widehat{PA}_w(0,l)$, defined in Definition 3, where all w-blocking arrivals (i.e., the elements of D_1) are diverted in the first step, before the thresholding is applied in the second step. However, because we are now concerned with the case where w is finite, we will make a further differentiation among the elements in D_1, by using another (equivalent) version of the $\widehat{PA}_w(0,l)$ policy, given as follows.

Definition 4 *Consider an alternative version of the $\widehat{PA}_w(0,l)$ policy, consisting of the following steps.*

1. *The first step is divided into two sub-steps:*

   (a) *Apply the policy $PA_\infty(0,\infty)$ to the baseline queue length process, Q^0. Denote by Q_{PA_∞} the resulting queue length process.*

   (b) *Apply the policy $PA_w(0,\infty)$ to the queue length process Q_{PA_∞}. Denote by Q_{PA_w} the resulting queue length process, and by A_{PA_w} the remaining arrivals in A that are not diverted in either 1.a or 1.b.*

2. *Apply an online threshold policy, with threshold l, to a system with arrival process A_{PA_w}, service token process S, and an initially empty queue.*

Let D_1^i be the set of elements of D_1 which are also ∞-blocking, and let $D_1^f = D_1 \backslash D_1^i$ be its complement. This partitioning of D_1 is done so that D_1^i corresponds to the set of diversions that would have been made under any length of the lookahead window, while the composition of D_1^f depends on the precise value of w.

It is not difficult to verify that the diversions made in Steps 1.a and 1.b correspond to D_1^i and D_1^f, respectively. Therefore, the set of all diversions made during Steps 1.a and 1.b, D_1, coincides with those made during Step 1 in the original version of the $\widehat{PA}_w(0,l)$ policy, in Definition 3, and the equivalence between $\widehat{PA}_w(0,l)$ and $PA_w(0,l)$ (Lemma 1) continues to hold for this alternative version of $\widehat{PA}_w(0,l)$ as well. We

will therefore focus on characterizing the diversion rate resulted from the $\widehat{PA}_w(0,l)$ policy in Definition 4 for the remainder of the proof.

Denote by d_1^i and d_1^f the rates of D_1^i and D_1^f, respectively. The value of d_1^i is easy to compute: from Eq. (16) in the proof of Theorem 1, we have that

$$d_1^i = \lambda - (1-p). \tag{29}$$

Our main task will be to compute d_1^f, which depends on the value of w. Denote by A_{PA_∞} the counting process associated with the arrivals in Q_{PA_∞} (i.e., the set of upward jumps in Q_{AP_∞}), and let t_k be the time of the kth arrival in A_{AP_∞}. Define

$$R_k = \inf\{s \in \mathbb{R}_+ : Q_{PA_\infty}(t_k + s) \leq Q_{PA_\infty}(t_k^+) - 1\}. \tag{30}$$

Recalling the definition of w-blocking arrivals (Definition 1), we see that the kth arrival in Q_{PA_∞} belongs to the set D_1^f if and only if $R_k \geq w$. By Proposition 1, Q_{AP_∞} is distributed according to the queue length process associated with an $M/M/1$ queue with arrival rate $1-p$ and service rate λ. Therefore, for any fixed $k \in \mathbb{N}$, we have that R_k is distributed according to the busy period of an $M/M/1$ queue with the same parameters. Using standard results on the probability density function of the busy period in an $M/M/1$ queue (cf. Chapter 2, Gross et al. (2013)), we have that

$$\mathbb{P}(t_k \in D_1^f) = \mathbb{P}(R_k \geq w) = 1 - F_{1-p,\lambda}(w) = \int_w^\infty \frac{1}{s\sqrt{\beta}} e^{(1-p+\lambda)s} I_1(2s\sqrt{\lambda(1-p)})\, ds, \quad \forall k \in \mathbb{N}, \tag{31}$$

where the notation $t \in D_1^f$ means that there is a diversion in D_1^f occurring at time t. We further observe that, because Q_{PA_∞} is distributed according to the queue length process of an $M/M/1$ queue, the arrival process A_{PA_∞} is Poisson with rate $1-p$, and, in particular, that

$$\lim_{t\to\infty} \frac{1}{t} A_{PA_\infty}(t) = 1-p, \quad a.s. \tag{32}$$

We have that

$$\begin{aligned} d_1^f &= \limsup_{t\to\infty} \frac{1}{t}\mathbb{E}\left(D_1^f(t)\right) = \limsup_{t\to\infty} \frac{1}{t}\mathbb{E}\left(\sum_{k=1}^{A_{PA_\infty}(t)} \mathbb{I}(t_k \in D_1^f)\right) \\ &\overset{(a)}{=} \lim_{t\to\infty} \frac{1}{t} \sum_{k=1}^{(1-p)t} \mathbb{P}(t_k \in D_1^f) \\ &\overset{(b)}{=} (1-p)(1 - F_{1-p,\lambda}(w)), \end{aligned} \tag{33}$$

where step (a) follows from Eq. (32) and the fact that $\mathbb{I}(t_k \in D_1^f) \leq 1$, and (b) from Eq. (31).

Finally, denote by d_2 the diversion rate induced by Step 2 of the $\widehat{PA}_w(0,l)$ policy in Definition 4. Because the set of diversions made during Steps 1.a and 1.b when $w < \infty$ is almost surely no smaller than that under $w = \infty$, it is not difficult to show, via a coupling argument, that the set of diversions made during Step 2 in our scenario is almost surely no greater than those made when $w = \infty$. Therefore, we have that for all $w < \infty$,

$$d_2 \leq (1-p)\frac{1-\beta}{1-\beta^{l+1}}\beta^l, \tag{34}$$

with $\beta = \frac{1-p}{\lambda}$, where the right-hand side corresponds to the value of d_2 when $w = \infty$ (cf. Eq. (17)).

Combining the diversion rates from all steps of $\widehat{PA}_w(0,l)$, Eqs. (29), (33), and (34), yields the upper bound

$$\begin{aligned} r_{w,l} =& d_1^i + d_1^f + d_2 \\ \leq& \lambda - (1-p) + (1-p)\left(1 - F_{1-p,\lambda}(w)\right) + (1-p)\frac{1-\beta}{1-\beta^{l+1}}\beta^l \\ =& \lambda - (1-p)\frac{1-\beta^l}{1-\beta^{l+1}} + (1-p)(1 - F_{1-p,\lambda}(w)), \end{aligned} \tag{35}$$

and lower bound

$$r_{w,l} \geq d_1^i + d_1^f = \lambda - (1-p) + (1-p)(1 - F_{1-p,\lambda}(w)) = \lambda - (1-p)F_{1-p,\lambda}(w). \tag{36}$$

This completes the proof of Theorem 3. □

A.4. Proof of Lemma 1

Proof of Lemma 1: For simplicity of notation, for the remainder of the proof of Lemma 1 we shall write PA_w and $\widehat{PA}_w$ in place of $PA_w(0,l)$ and $\widehat{PA}_w(0,l)$, respectively. We will represent a point process as a set of points in $\mathbb{R}_+$. For instance, if D is the point process that corresponds to the set of diversions, then the notation $t \in D$ means that there is an arrival being diverted at time t.

Denote by $\widehat{D}_1$ and $\widehat{D}_2$ the sets of diversions made during the first and second step of $\widehat{PA}_w$, respectively, so that $\widehat{D} = \widehat{D}_1 \cup \widehat{D}_2$. By definition, $\widehat{D}_1$ is equal to the set of all w-blocking arrivals in A, and because a w-blocking arrival must be diverted under PA_w, we have that $\widehat{D}_1 \subset D$. Let $D_2 = D \backslash \widehat{D}_1$ be the set of diversions made by PA_w that is not w-blocking. It remains to show that

$$D_2 = \widehat{D}_2. \tag{37}$$

Let $\{Q(t)\}_{t\in\mathbb{R}_+}$ and $\{\widehat{Q}(t)\}_{t\in\mathbb{R}_+}$ be the queue length processes induced by PA_w and $\widehat{PA}_w$, respectively. Let $\Delta = (D_2 \backslash \widehat{D}_2) \cup (\widehat{D}_2 \backslash D_2)$ be the symmetric difference between D_2 and $\widehat{D}_2$. Let $t^* = \inf \Delta$, and, for the sake of contradiction, suppose that $t^* > -\infty$, for otherwise $D_2 = \widehat{D}_2$.

Suppose $t^* \in D_2$. By the definition of D_2, the arrival at t^* must not be w-blocking, and hence we have that $Q(t^{*-}) = l$. Meanwhile, the definition of t^* implies that the sets D and $\widehat{D}$ agree up to time t^*, i.e., $D \cap [0, t^*) = \widehat{D} \cap [0, t^*)$, where we use the notation $D \cap [0,t)$ to denote the set of all points in D before time t. This implies that $\widehat{Q}(t^{*-}) = Q(t^{*-}) = l$, which means the arrival at time t^* will also be diverted by the second step of $\widehat{PA}_w$, and hence $t^* \in \widehat{D}_2$, which contradicts with the definition of t^*. The above arguments show that we must have $t^* \in \widehat{D}_2$. We can apply an identical argument to show that $t^* \in \widehat{D}_2$ would imply that t^* also belongs to D_2, leading to a contradiction with the assumption that $t^* > -\infty$. We thus conclude that $D_2 = \widehat{D}_2$, which proves our claim. □

B. Other Materials

B.1. Connections Between the No-Show Noise Model and a Noisy Predictive Model for Arrival Counts

B.1.1. Coefficient of Determination We show how the parameter ϵ for the level of no-show noise is related to a predictive model's coefficient of determination, R^2. Consider a small time interval, $h = [t, t+\delta)$,

and denote by X_h and X'_h the realized and predicted numbers of arrivals during h, respectively. Under the no-show noise model, we have that the difference $(X'_h - X_h)$ is distributed as a Poisson random variable with mean and variance $\delta\lambda\frac{\epsilon}{1-\epsilon}$. Since the variance of the true arrival count, X_h, is equal to $\delta\lambda$, we have that the coefficient of determination for the arrival count prediction during the interval h is given by

$$R_h^2 = 1 - \frac{\text{var}(X'_h - X_h)}{\text{var}(X_h)} = 1 - \frac{\delta\lambda\epsilon}{\delta\lambda(1-\epsilon)} = \frac{1-2\epsilon}{1-\epsilon}. \tag{38}$$

Note that the value of R_h^2 is independent of the length of the interval h. Therefore, the no-show noise model with parameter ϵ corresponds to a noisy predictive model for arrival counts, whose coefficient of determination is equal to $(1-2\epsilon)/(1-\epsilon)$ for predicting the total arrival counts over any finite time interval, and whose predictions over disjoint time intervals are independent.

B.1.2. Mean Absolute Percentage Error We now show how the parameter ϵ for the level of no-show noise is related to the Mean Absolute Percentage Error (MAPE) of a predictive model. Define X_h and X'_h as before. Under the no-show noise model, we have that the difference $X'_h - X_h \geq 0$ and has mean $\delta\lambda\frac{\epsilon}{1-\epsilon}$. Since the mean of the true arrival count, X_h, is equal to $\delta\lambda$, we have that the MAPE for the arrival count prediction during the interval h is given by

$$\text{MAPE}_h = \frac{\mathbb{E}[|X_h - X'_h|]}{\mathbb{E}[X_h]} = \frac{\epsilon}{1-\epsilon}. \tag{39}$$

Again, note that the value of the MAPE_h is independent of the length of the interval h.

B.2. Proactive Policy for Systems with Priority Arrivals

The w-proactive policies proposed in Definition 2 can also be adapted to work in a scenario where a subset of the arrivals is given *priority* and is in general not to be diverted. In the ED setting, this can correspond to having high-acuity patients who are often admitted for treatment immediately. For concreteness, consider a system where the arrival process, A, is the superposition of two independent Poisson processes: A_p, with rate $\lambda_p < 1-p$, which corresponds to the priority jobs, and A_o, with rate $\lambda - \lambda_p$, which corresponds to the ordinary arrivals. The arrivals in A_p are always admitted and wait in a *priority queue*. A subset of the jobs in A_o can be admitted and wait in an *ordinary queue*, with the remainder of A_o being diverted. The service tokens in S are always used to first serve jobs in the priority queue, and will be used to serve jobs in the ordinary queue only when the priority queue is empty. We use this strict priority scheduling rule for illustrative purposes, and other rules can be incorporated analogously.

To generate the set of diversions in this system using a proactive policy, we will take into account the impact on service availability due to the priority arrivals, as follows. Let $S_o \subset S$ be the set of service tokens that are generated when the priority queue is empty, or, equivalently, the set of service tokens that can be used to serve ordinary arrivals. We will simply apply the same w-proactive diversion policy (Definition 2) to a system with arrival process A_o and service token process S_o.

Note that the only difference from the original model is that S_o is no longer a Poisson process, but a Markov-modulated Poisson process (MMPP) with rate $(1-p)-\lambda_p$, which has an instantaneous rate of $1-p$

when the priority queue is empty, and 0 otherwise. Our analytical results do not directly apply to this more general class of service processes. However, when value of λ_p is not too large, the priority queue visits the empty state often. In Peck et al. (2012), the most severe triage class (ESI 1) comprises less than .2% of the patients who visit the ED, while the two most severe patient classes (typically those considered to not be divertible) comprise less than 2% of ED arrivals. In such settings, the MMPP service token process is statistically similar to that of a Poisson process with the same rate, and it is reasonable expect that the resulting performance will be close to that of the original proactive diversion policy.