Models of the Impact of Triage Nurse Standing Orders on Emergency Department Length of Stay

Saied Samiedaluie
Alberta School of Business, University of Alberta
samiedal@ualberta.ca

Vera Tilson
Simon Business School, University of Rochester
vera.tilson@simon.rochester.edu

Armann Ingolfsson
Alberta School of Business, University of Alberta
armann.ingolfsson@ualberta.ca

Problem Definition: Standing orders allow triage nurses in emergency departments (EDs) to order tests for target patients prior to a physician evaluation. Standing orders specify the medical conditions for which a triage nurse is permitted to order tests but typically do not specify the operational conditions under which ordering tests is desirable, from either a system or a patient point of view. **Academic/Practical Relevance:** Medical studies demonstrate that the use of standing orders decreases average ED length of stay (LOS) for target patients. We examine the operational impacts of standing orders on the ED as a whole, and propose a threshold policy for enacting standing orders as a function of ED congestion. **Methodology:** We develop three simplified models: 1) an infinite-server model to study how model primitives impact the effectiveness of standing orders, 2) a Jackson network model, to demonstrate that standing orders can lead to diverse outcomes for different patient populations, and 3) a Markov decision process model, to quantify the optimality gap for our threshold policy. We confirm the tentative findings from the simplified models in a more realistic setting using a simulation model that is calibrated with real data. **Results:** We find that the threshold policy, with a threshold that is easily estimated from model primitives, performs well across a wide range of parameter values. We demonstrate potential unintended consequences of the use of standing orders, including overtesting and spillover effects on non-target patients. **Managerial Implications:** Our research shows the importance of investigating the impact of standing orders on the ED as a whole. The simple threshold policy that we propose can be used in practice to identify situations in which it is beneficial to use standing orders.

1. Introduction

Chest pain is the second most common symptom of patients presenting to emergency departments (EDs) (Rui and Kang 2017). Since chest pain can indicate a life-threatening condition, medical guidelines recommend an electrocardiogram (ECG) within 10 minutes of arrival to an ED (Zègre-Hemsey et al. 2016). *Standing orders* (also termed *complaint specific protocol*, or *advanced triage protocol*) are one way EDs can ensure that the ECG is performed in a timely fashion. Triage healthcare providers not explicitly licensed to

place orders for medical tests and treatments can initiate an ECG by following a codified protocol for patients presenting with chest pain.

Standing orders can improve outcomes for patients with life-threatening emergencies, but the practice need not be limited to critical cases. For example, a patient presenting to ED with a low acuity complaint, such as sore throat, is likely to be triaged to low priority and have to wait a long time for an initial evaluation by a physician. Following the standing order protocol in Figure 1, a time-consuming bacterial culture test can be ordered at triage and processed while the patient is waiting for a physician. For the target patients subject to the sore throat protocol, this process intervention can reduce the ED length of stay (LOS) by reordering the customary ED process flow (Settelmeyer 2018).

A typical ED process flow (Figure 2) involves triage by a nurse, initial evaluation by a physician, diagnostic tests or treatments ordered by a physician, and the physician's disposition decision: discharge or hospital admission. This series of steps, with potential waiting periods between steps, constitutes the critical path, whose duration determines a patient's ED LOS, which is a key performance measure (Wiler et al. 2015). Standing orders can reduce ED LOS for target patients for two reasons: (1) Completing the tests during the time the patient waits for the initial physician evaluation moves the testing off the critical path, and (2) Having test results available in time for the patient's initial evaluation by a physician can eliminate the need for the physician to revisit the case.

Despite the potential benefits of ED standing orders, the practice is not universal. Reported obstacles to adoption include: increased workload for triage nurses, perceived lack of operational benefits if triage-ordered test results are not available in time for the initial physician evaluation, inability of some EDs to reliably change their processes, and costs of *overtesting* (Retezar et al. 2011). Furthermore, the practice of using mid-level providers to order tests at triage has been criticized as prioritizing hospital profits over patient care and leading to overtesting (Corl 2019). Overtesting incurs not only the direct cost of performing and interpreting a test, but also indirect costs, such as increased load on testing resources, leading to congestion, and lengthening the wait for *all* patients who need testing—not only the target patients for standing orders. Overtesting can also increase the load on physicians who must review the results of unnecessary tests to minimize the chances of missing an incidental finding, which can increase ED wait times even more.

EDs have sought to understand when and how to use standing orders to reduce ED congestion. However, a comprehensive review of empirical studies in the medical literature

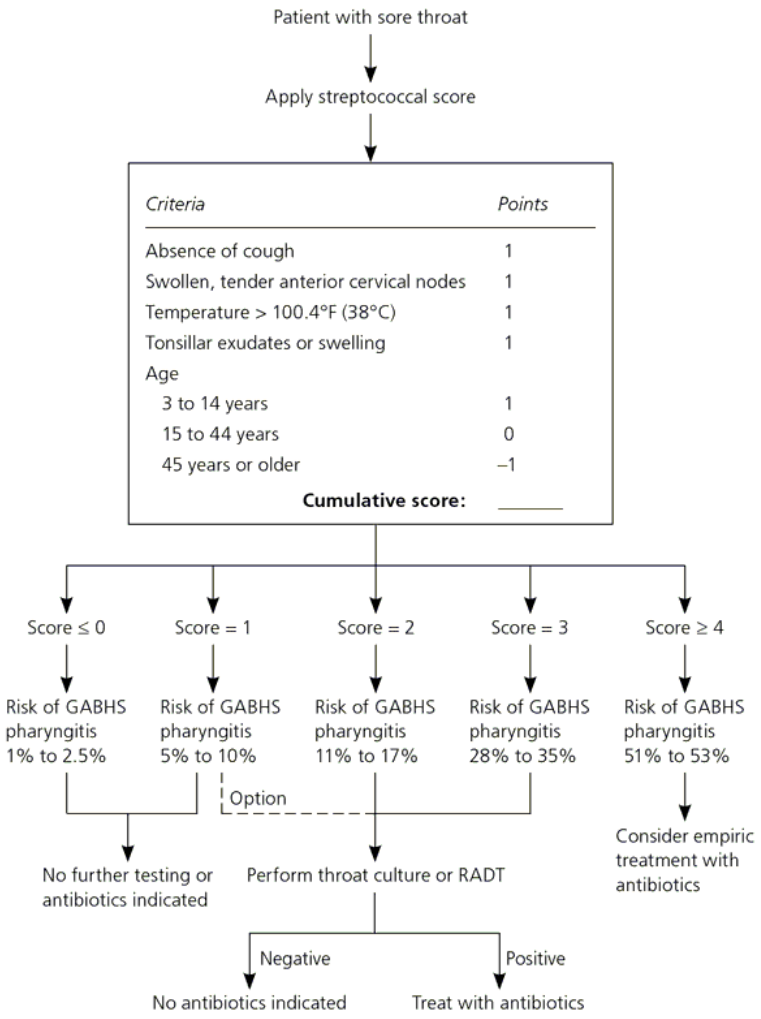

Figure 1 **Rapid strep test protocol. Source: Kalra et al. (2016).**

reveals that the overwhelming majority solely examine the effect of the intervention on the target population (patients presenting with specific symptoms that are subject to the standing orders protocol). One important contribution of our research is the demonstration of the necessity of taking a system-wide perspective. With numerical experiments using operational parameters reported in the medical literature, along with simulation models that are calibrated with real ED data, we demonstrate potential unintended consequences of standing orders, including spillover effects on ED patients who are not subject to the standing order protocol. We also show that for a realistic set of parameters, an increase in ED LOS for the target population can be accompanied by a decrease in the overall average ED LOS, as non-target patients benefit from reduced wait for service by an ED physician. Thus, empirical researchers should measure the effect of the intervention both on target and non-target patients.

Initiation of standing orders is an example of *early task initiation* (ETI), an important mechanism through which system load can influence service time in a queueing network (Batt and Terwiesch 2017, Delasay et al. 2016). Observational studies have found that the use of ETI increases with congestion (Batt and Terwiesch 2017). In addition to studying how activation of ETI by human servers depends on congestion and other factors, it is also important to study optimal or near-optimal policies for using ETI to reduce ED LOS (Batt and Terwiesch 2017), and that is another area where we make an important contribution.

We consider not only the two extreme regimes: never invoking standing orders (NSO) and always invoking standing orders (ASO), but also a dynamic use of standing orders based on operational conditions in the ED. Our review of empirical medical literature identified only three papers discussing operational conditions under which ordering of tests at triage is *desirable*: Hwang et al. (2016) and Retezar et al. (2011) specify that standing orders should be initiated when the time until initial evaluation by a physician is expected to be longer than some threshold and Li et al. (2018) recommend invoking standing orders when the time until test results are available is below a threshold. Reliable estimates of processing times or delays are not easy to obtain by the triage nurse. Information about the number of patients waiting in different parts of the ED is likely easier to obtain, and this is the type of information that we assume is available for the policy we propose: initiate standing orders when the difference between the number of patients waiting for a physician exceeds the number of patients waiting for testing by a certain threshold (see (1)). This threshold is easily computed from model primitives: target population percentage, overtesting rate, and mean service times of an ED physician and the testing station.

Our work appears to represent the first suite of analytical models of ED standing orders. We quantify the relationship between target population percentage, overtesting rate, and mean service times of the ED physician and the testing station to identify whether the overall ED LOS improves with the use of standing orders. Viewing an ED as a queueing network, we develop three simplified analytical models: 1) an infinite-server model to study how model primitives impact the effectiveness of standing orders, 2) a Jackson network model, to demonstrate that standing orders can lead to diverse outcomes for different patient populations, and 3) a Markov decision process (MDP) model, to characterize congestion-sensitive routing policies that minimize the overall average ED LOS. We demonstrate with extensive MDP numerical experiments that both (1) an *optimal threshold policy*, obtained through complete enumeration, and (2) an *approximate threshold*

policy, for which the threshold is expressed as a surprisingly simple function of the model primitives, are near-optimal across a wide range of parameters. We use a discrete event simulation (DES) model to confirm that the approximate threshold policy performs well in a more realistic setting, incorporating multiple servers, non-exponential distributions, and a non-stationary arrival process.

The rest of the paper is organized as follows: Section 2 reviews related literature; Section 3 describes a typical ED process flow and how the use of standing orders changes the flow; Section 4 presents our three analytical models; and Sections 5-6 report results of numerical experiments on the MDP and DES models. Section 7 concludes.

2. Literature Review

We review four streams of literature relevant to our study: (1) medical studies on the impact of standing orders on ED LOS, (2) DES studies of standing orders, (3) studies of staffing triage with providers licensed to place medical orders (an alternative to standing orders), and (4) OM papers examining costs and benefits of various alternatives for post-triage prioritization of ED patients.

Stream 1: Table 1 summarizes findings from 17 medical studies of standing orders. Nine studies were included in a 2011 systematic review (Rowe et al. 2011), and eight were published after 2011. Eleven studies investigated standing orders for ordering X-rays for limb injuries. The majority of these are studies of the *Ottawa Ankle Rules* protocol to determine whether a patient presenting with an ankle injury requires an X-ray (Shell et al. 1993). The remaining studies examined standing orders for patient complaints such as throat pain, pediatric emergency, and chest pain.

The primary outcome measure for most studies is ED LOS for the target population and most studies report reductions in this measure, ranging in magnitude from 4 to 46 minutes (4.3 to 36%), as shown in Table 1 (Column 4). (See Table 11 in Appendix A for a summary of results from studies whose outcome measures were defined differently, in terms of start point, end point, or patient inclusion criteria). A single study (Hwang et al. 2016) found an increase in ED LOS for the target population but this study failed to control for ED congestion levels—standing orders were used only when the ED was congested, when patients were experiencing longer waiting times.

Several other factors were identified as important to the operational effectiveness of standing orders. If triage nurses order more tests than physicians, the potential benefits of standing orders can be lost (Thurston and Field 1996). Having the test results ready

before the physician sees the patient is another critical factor. In Parris et al. (1997), where no significant LOS reduction was found, only 77% of patients had their X-rays ready before being evaluated by the physician. Hwang et al. (2016) report that having the tests completed reduced the time from physician evaluation to disposition by 16.9%.

With the exception of Thurston and Field (1996) and Rosmulder et al. (2010), all of the empirical studies focus only on the effect of ASO on target patients, and do not discuss the effect on the system overall (see Table 11). Our study demonstrates that a reduction in the LOS for the target patients is neither necessary nor sufficient for an improvement in the overall system performance: Shorter target-patient LOS can be accompanied by longer overall LOS, because of overtesting, and longer target-patient LOS can be accompanied by shorter overall LOS, because of freeing up of physician capacity.

Stream 2: We know of two DES studies (Ghanes et al. 2015, Yang et al. 2016) that investigate factors (listed in Table 2) that impact the effectiveness of using standing orders to reduce overall average ED LOS. Both studies assume that standing orders are used for all target patients, rather than congestion-triggered policies. Our paper expands the OM literature on standing orders by using a suite of analytical models to generate insight and using a DES model to evaluate congestion-triggered protocols.

Stream 3: One reason for limited adoption of standing orders is regulation: ED triage is generally staffed with registered nurses, who are allowed to use standing orders in some US states (e.g., Colorado), but not in others (e.g., New York) (Castner et al. 2013). In locations where the use of standing orders by registered nurses is not allowed, staffing triage with a physician or a similarly licensed provider is an alternative to standing orders (Russ et al. 2010, Nestler et al. 2012, Kamali et al. 2018). This practice introduces a trade-off between placing such providers at triage or at later stages in the ED process flow. At least one medical study has found that staffing triage with physician assistants reduced both rates of leaving without being seen (LWBS) and the time patients spend occupying an ED bed (Nestler et al. 2012). An empirical study of *physician-staffed triage* also reported a decrease in the time patients spent occupying ED beds; however the median time from arrival to disposition *increased* (Russ et al. 2010). OM scholars have investigated policies for the allocation of physicians to triage vs. post-triage stages, so as to optimize time to first treatment and timely discharges (Zayas-Caban et al. 2019) or to optimize the trade-off between staffing cost and revenue loss from patients who leave without being seen (Kamali et al. 2018). Our study complements this work, as the threshold policies that we propose

Target population, test type	Congestion triggered	Study design	LOS reduction for target population	Sample size	Reference
Limb injury, X-rays	No	RCT	4 min (4.3%)[a]	1,833	Thurston and Field (1996)
			NR	175	Parris et al. (1997)
			37.2 min (36%)***	675	Lindley-Jones and Finlayson (2000)
			6.7 min (8.4%)[a]	130	Fan and Woolfrey (2006)
			28 min (19.6%)**	146	Lee et al. (2016)
			13 min (14.9%)*	112	Ho et al. (2018)
		PC	NR	934	Lee et al. (1996)
				106	Pedersen and Storm (2009)
		B-A	14 min (14%)	704	Rosmulder et al. (2010)
			6.5 min (6.3%)[a]	60	Ashurst et al. (2014)
Limb/skull injuries, X-rays	No	PC-C	24.5 min	276	Than et al. (1999)
Chest pain, Multiple	Yes	RC	-212 min (-52.7%)***	301	Hwang et al. (2016)
Pediatric emergency, Multiple	Yes	RC	15 min (6.2%)***	116,202	Li et al. (2018)
Throat pain, Multiple	No	RC	6 min	117	Settelmeyer (2018)
Multiple, Multiple	No	B-A	46 min	250	Cheung et al. (2002)
	Yes	RC-C	NR	15,188	Retezar et al. (2011)
	No	RCT	NR	1,044	Goldstein et al. (2018)

Table 1 **Summary of medical literature findings on the impact of standing orders initiated by triage nurse.**
Legend: NR = not reported, RCT = randomized controlled trial, B-A = before-after study, PC = prospective cohort study, PC-C = prospective case-controlled study, RC = retrospective cohort study, RC-C = retrospective case-controlled study.
Legend for statistical significance of LOS reduction: *: $p < 0.05$**, **:** $p < 0.01$**, ***:** $p < 0.001$**, a: not statistically significant, no code: statistical significance not reported.**

Reference	Parameters	Standing orders effective when:
Ghanes et al. (2015)	ED load	ED load is high
	Extension of triage time due to standing orders	Time extension is reasonable
	Overtesting and incomplete test rates	Inaccuracy is low
Yang et al. (2016)	Physician utilization	Physician utilization is high
	Triage and test capacities	Not important

Table 2 **DES studies of triage standing orders.**

could be used to guide licensed providers regarding conditions under which ordering tests early would be beneficial.

Stream 4: OM researchers have modeled other triage routing possibilities besides early ordering of tests via either standing orders or physician-staffed triage. In particular, they have investigated *streaming* of patients based on acuity level (Cochran and Roche 2009), predicted disposition (discharged or admitted to hospital) (Saghafian et al. 2012), or predicted complexity (number of patient interactions in the ED) (Saghafian et al. 2014). OM researchers have also examined prioritization by ED physicians of newly arrived patients vs. in-process patients. The contexts include in-process patients creating additional work for a physician through interruptions (Dobson et al. 2013), considerations of trade-offs between the time-to-first treatment and LOS (Huang et al. 2015, Hu et al. 2018), and lack of information about the state of ED queues (Ansari et al. 2019). Other OM papers that use queueing theory to study management of EDs are reviewed by Hu et al. (2018). Our study expands this stream, by investigating a new set of ED patient routing issues.

3. ED Process Flow, with and without Standing Orders

In this section, we describe a typical ED process flow and discuss how the implementation of standing orders changes the flow.

3.1 ED Process Flow without Standing Orders

Figure 2 illustrates a typical ED process flow without standing orders. Patients register upon arrival and wait for triage. Triage is performed by a nurse who determines the patient's primary reason for visiting the ED, and assesses how urgently the patient needs to be evaluated by an ED physician. In the order of priority determined by the triage nurse, the patient is allocated an ED bed, and a physician performs an initial evaluation. In some cases, the physician treats and discharges the patient during this initial exam (e.g., a patient needs stitches for a cut). In other cases, the physician requires additional information to arrive at a diagnosis and reach a decision on patient disposition. The additional information may come from diagnostic tests, or from responses to treatment.

A diagnostic test could be performed at the bedside in the ED (e.g., a pregnancy ultrasound), in a hospital lab using a sample collected in the ED (e.g., a complete blood count), or at another hospital unit (e.g., an MRI scan). Depending on the availability of resource needed for testing/treatment, a patient may have to wait to be tested or wait to have a treatment started. There can be a significant delay between the last step in the test or treatment process which involves the patient (e.g., drawing a blood sample) and the

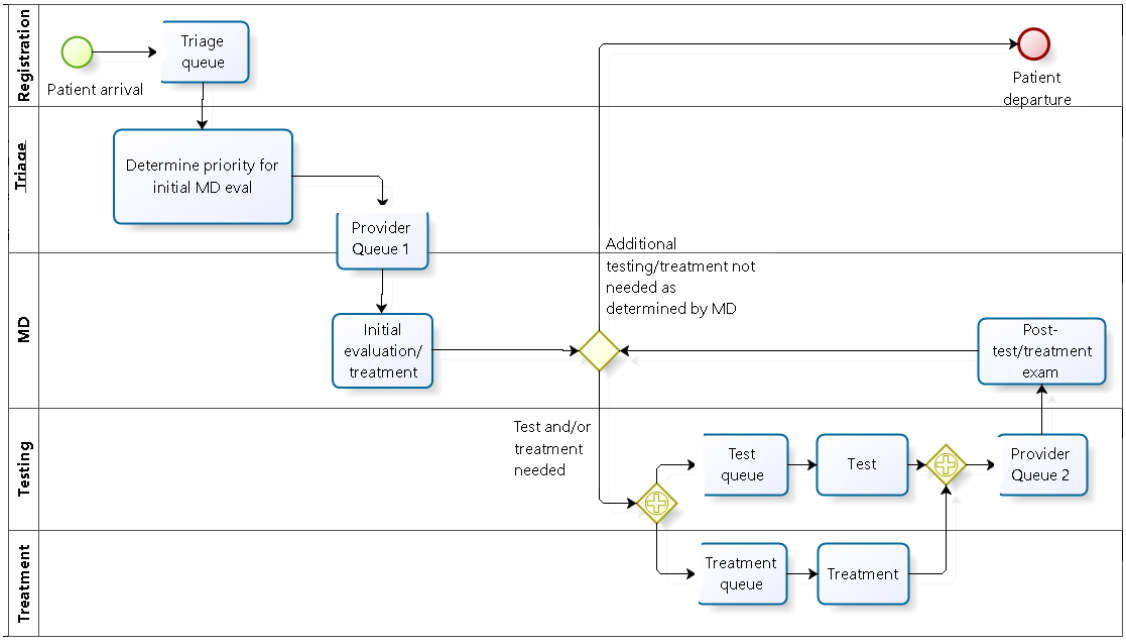

Figure 2 **The default ED care pathway.**

instant when the test results or treatment response become available. After this information becomes available, the ED physician decides whether to discharge or to admit the patient to a hospital ward. The patients subsequently depart the ED, although a patient directed to a hospital ward may need to wait in an ED bed until a ward bed becomes available.

Patients might abandon, either prior to initial evaluation ("left without/before being seen") or after initial evaluation but prior to disposition ("left against medical advice"). We omit abandonment from Figure 2 and from our models, in order to focus on process flow features that are most relevant to our study.

3.2 ED Process Flow with Standing Orders

Figure 3 depicts an ED process flow with standing orders. The triage nurse decides whether the patient's condition is covered by a standing orders protocol. If not, then the process continues as before. If the patient is deemed to be in the target group, then there is a second decision: whether to invoke the test ordering protocol. Placing test and treatment orders increases the nurse's workload. Normal practice appears to be to use standing orders whenever applicable, without consideration of operational factors such as ED congestion, but see Hwang et al. (2016), Retezar et al. (2011), and Li et al. (2018) for exceptions.

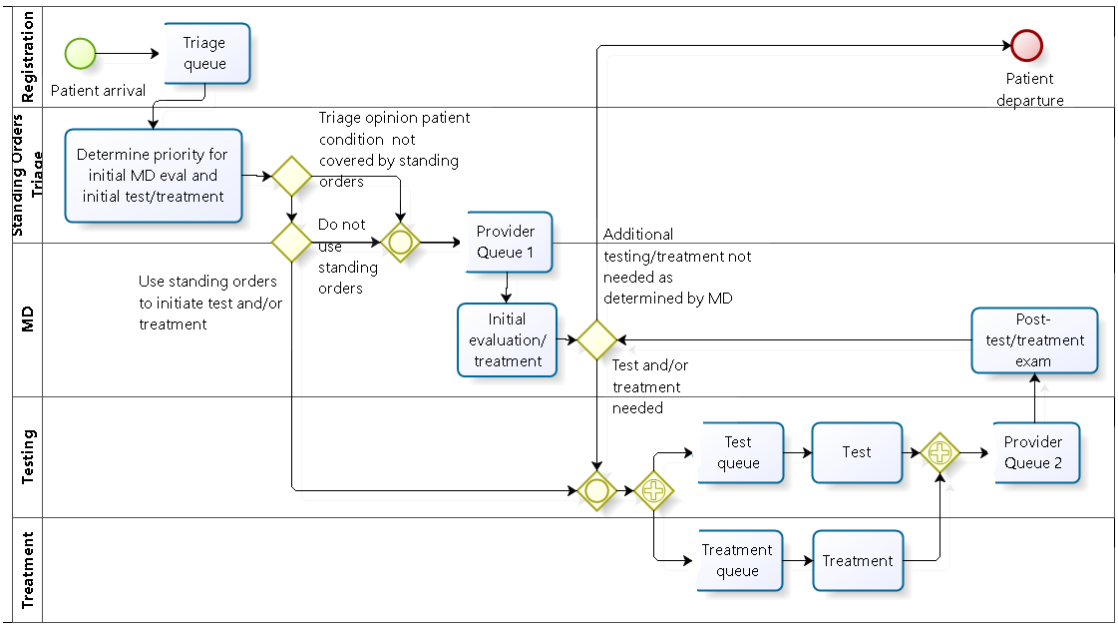

Figure 3 **ED care pathway with standing orders.**

4. Analysis of Standing Order Routing Policies

The focus of this section is on generating insights regarding policies for post-triage routing of patients who are identified at triage as presenting with a medical condition that is subject to a standing order. The routing policies specify operational conditions under which patients should be routed from triage directly to testing. We are particularly interested in the performance of a policy that computes the difference between the total number of patients queued for or receiving physician service (Q_{phys}) and the number of patients queued for or receiving testing (Q_{test}), and routes patients to testing if this difference is larger than a threshold, that is:

$$\begin{aligned} &\text{If } Q_{\text{phys}} - Q_{\text{test}} \geq \theta, \text{ then route to testing,} \\ &\text{Otherwise, route to a physician.} \end{aligned} \tag{1}$$

This is a *threshold policy* with a single fixed threshold θ. The ASO and NSO policies are special cases of (1), corresponding to $\theta = -\infty$ and $\theta = +\infty$, respectively.

We discuss four models of ED flow with standing orders: An infinite-server model, a Jackson network model, an MDP model, and a DES model. The first three models are highly simplified but they allow us to generate tentative insights. We use the last model to confirm that these insights hold in a more realistic setting.

In all models, we consider only the time period that begins with a patient's placement in a treatment space and finishes when a disposition decision is made. Patients do not

Probabilities	Predicted: Target	Predicted: Other	
True: Target	$\eta_{TT} = \psi$	$\eta_{TO} = 0$	$\eta_{T*} = \psi$
True: Other	$\eta_{OT} = (1-\nu)(1-\psi)$	$\eta_{OO} = \nu(1-\psi)$	$\eta_{O*} = 1-\psi$
	$\eta_{*T} = 1-\nu(1-\psi)$	$\eta_{*O} = \nu(1-\psi)$	$\eta_{**} = 1$

Table 3 Distribution of patient types

abandon, they have at most one test encounter, and only one post-test evaluation with an ED physician. Test results are available as soon as the testing service completes.

We view the ED as a multi-class queuing network. A patient's class can depend both on their true type, as determined by an ED physician, and on their predicted type, as determined by a triage nurse. We assume that the probability that a physician determines that the patient belongs to the standing orders *target* population and therefore needs testing is ψ. We are particularly interested in the effect of overtesting, so we assume that the triage nurse correctly identifies *other* patients with probability ν (and therefore, under ASO, other patients get tested with probability $1-\nu$). We assume that the triage nurse correctly identifies all target patients. We use a two-letter notation to indicate patient types, with the first letter indicating the true type and the second letter indicating the predicted type. We let $T =$ target, $O =$ other, and $* =$ all. Table 3 shows the joint and marginal probabilities η_p for $p \in P = \{TT, TO, T*, OT, OO, O*, *T, *O, **\}$.

Comparing Figures 2–3, we see that NSO and ASO induce different queueing network topologies. Under the assumption of at most one test and post-test evaluation, the primary difference is that the NSO network has re-entrant flows, which complicates modeling, but the ASO network does not. For the purposes of the infinite-server and Jackson network models, we make the simplifying assumption that there are separate physician resources dedicated to initial evaluation and post-test evaluation, respectively. The resulting topology is shown in Figure 4. This network has an initial evaluation physician queue Q_1, with mean processing time τ_1, a testing queue Q_2, with mean processing time τ_2, and a post-test evaluation physician queue Q_3, with mean processing time τ_3. We assume $\tau_i < \infty, i = 1, 2, 3$. In addition to eliminating the re-entrant flows, the simplifying assumption of dedicated servers for each queue also eliminates the need for a physician to choose between Queues 1 and 3 when selecting the next patient to see. For the MDP and DES models, we use the topology shown in Figure 5, which includes re-entrant flows, and separates patients waiting for initial evaluation into those triaged as target (waiting in Q_1 with mean processing time τ_1) and those triaged as other (waiting in Q_4 with mean processing time τ_4; we assume $\tau_1 = \tau_4$). For all of the models, we assume an independent Poisson arrival process for new

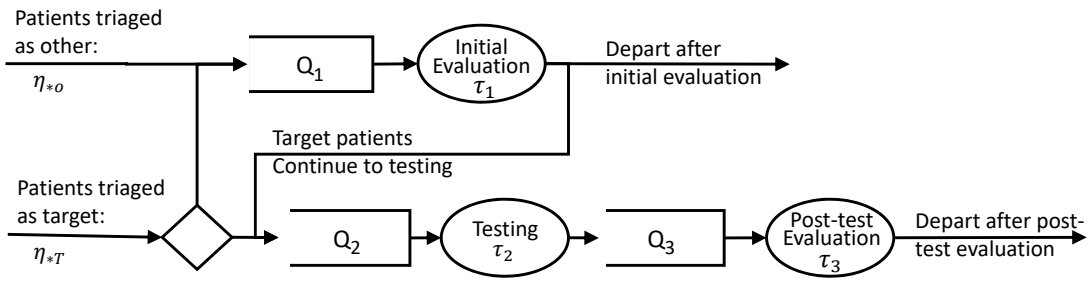

Figure 4 Queueing network topology for the infinite-server and Jackson network models

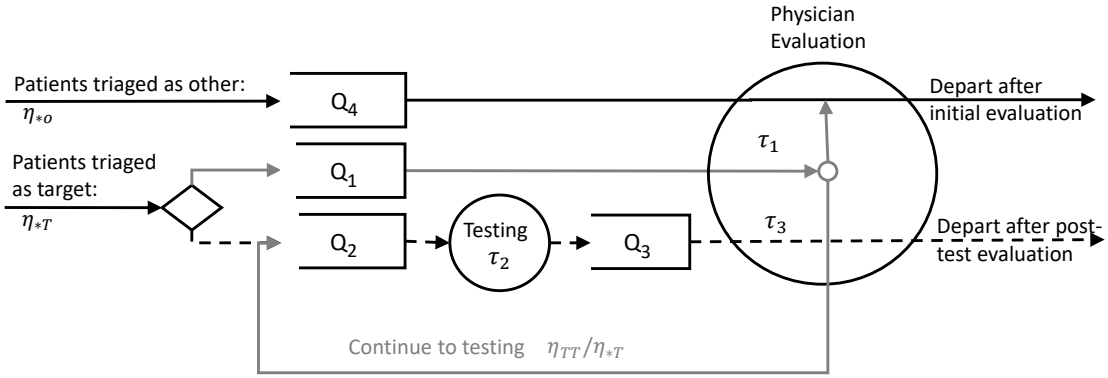

Figure 5 Queueing network topology for the MDP and DES models

patients, which are randomly split into patients triaged as target (with probability η_{*T}) and patients triaged as other (with probability η_{*O}). The Poisson process is homogeneous for the first three models but in the DES model, the process is nonhomogeneous.

We use $M = \infty$ and $M = J$ to denote the infinite-server model and the Jackson network model, respectively. Let $\text{LOS}_p^{\gamma,M}$ be the steady-state average LOS for patients of type $p \in P$ under policy $\gamma \in \{\text{NSO}, \text{ASO}\}$ derived using queueing model $M \in \{\infty, J\}$ and let $W_i^{\gamma,M}$ denote the average system time (waiting + processing) in queue Q_i under policy γ and model M. We define the change in LOS for patient type p under model M if standing orders are adopted as:

$$\Delta_p^M = \text{LOS}_p^{\text{ASO},M} - \text{LOS}_p^{\text{NSO},M},\ p \in P \tag{2}$$

$$\Delta_{**}^M = \sum_{p \in \{OO,OT,TT\}} \eta_p \Delta_p^M. \tag{3}$$

A positive Δ_p^M means that always using standing orders *increases* LOS for patients of type p, under model M. We determine Δ_{**}^M by substituting the η_p values from Table 3 and the following expressions for Δ_p^M into (3):

$$\Delta_{OO}^M = W_1^{\text{ASO},M} - W_1^{\text{NSO},M}, \tag{4}$$

$$\Delta_{OT}^{M} = W_2^{\mathrm{ASO},M} + W_3^{\mathrm{ASO},M} - W_1^{\mathrm{NSO},M}, \tag{5}$$

$$\Delta_{TT}^{M} = \Delta_{OT}^{M} - \left(W_2^{\mathrm{NSO},M} + W_3^{\mathrm{NSO},M}\right). \tag{6}$$

We obtain these expressions by tracing the path of OO, OT, and TT patient groups under ASO and NSO. Patients triaged as other (OO) follow the same path under both policies but the system time in Queue 1 could depend on the policy. Overtested patients (OT) go through Queues 2–3 under ASO and go only through Queue 1 under NSO. Target patients (TT) go through Queues 2 and 3 under ASO and go through all three queues under NSO.

We see from (5)–(6) that $\Delta_{OT}^{M} > \Delta_{TT}^{M}$, that is: if ASO lowers LOS for overtested patients, then it also lowers LOS for target patients. The reverse is not true, however: lower LOS under ASO for target patients does not imply lower LOS for overtested patients.

4.1 Infinite-Server Model

To focus on the effects of mean processing times on the relative performance of NSO and ASO, we formulate a network of infinite-server queues (Harrison and Lemoine 1981), by assuming the topology in Figure 4 and assuming that Queues 1–3 each have an infinite number of dedicated servers. We show that under this model, standing orders reduce overall LOS if the initial evaluation duration is long compared to testing and post-test evaluation durations (that is, $\tau_1/(\tau_2+\tau_3)$ is large) and both the target population proportion ψ and the triage accuracy ν are close to 1 (that is, $\psi/((1-\nu)(1-\psi))$ is large).

PROPOSITION 1. *Under the infinite-server model,* $\mathrm{LOS}_{**}^{\mathrm{ASO},\infty} < \mathrm{LOS}_{**}^{\mathrm{NSO},\infty}$ *if and only if*

$$\beta = \left(1 + \frac{\eta_{TT}}{\eta_{OT}}\right)\frac{\tau_1}{\tau_2+\tau_3} = \left(1 + \frac{\psi}{(1-\nu)(1-\psi)}\right)\frac{\tau_1}{\tau_2+\tau_3} > 1. \tag{7}$$

Proof: Under the infinite-server model, mean system time equals mean processing time. Substituting $W_i^{\gamma,\infty} = \tau_i$ into (4)–(6) and using (3) results in

$$\Delta_{**}^{\infty} = \mathrm{LOS}_{**}^{\mathrm{ASO},\infty} - \mathrm{LOS}_{**}^{\mathrm{NSO},\infty} = \eta_{OT}\left(\tau_2+\tau_3-\tau_1\right) - \eta_{TT}\tau_1. \tag{8}$$

Using (8), $\mathrm{LOS}_{**}^{\mathrm{ASO},\infty} < \mathrm{LOS}_{**}^{\mathrm{NSO},\infty}$ can be shown to be algebraically equivalent to $\beta > 1$. ■

As discussed in Sections 1–2, most empirical studies focus on the effect of ASO on target patients, and do not measure the effect on the system overall. In situations with plentiful resources and negligible queueing, ASO will shorten the LOS for target patients. But ASO can lengthen the LOS for overtested patients, and this can lengthen the overall LOS. As an example, consider the following extreme case: the test time is negligible ($\tau_2 \approx 0$), and all

arriving patients are tested (that is, ASO with $\nu = 0$). If the time spent reviewing unneeded test results increases τ_3 by a sufficient amount, specifically, if $\tau_3 > \tau_1/(1-\psi)$, then

$$\beta = \left(1 + \frac{\psi}{(1-\nu)(1-\psi)}\right)\frac{\tau_1}{\tau_2+\tau_3} \approx \left(1+\frac{\psi}{1-\psi}\right)\frac{\tau_1}{\tau_3} = \frac{\tau_1}{1-\psi}\frac{1}{\tau_3} < 1.$$

Thus, it can be preferable *not* to use standing orders even if the testing time is negligible.

We argue that β is predictive of the optimal θ^* for the threshold policy (1), as well as of which of the extreme policies, ASO or NSO, is preferable. To see this, note that β depends on two ratios: $\tau_1/(\tau_2+\tau_3)$ and η_{TT}/η_{OT}. A large $\tau_1/(\tau_2+\tau_3)$ means that the initial exam is long relative to the test and the post-test exam. A large η_{TT}/η_{OT} occurs if the proportion of target patients is high and the probability of overtesting is low. Both of these are situations in which one would expect that routing $*T$ patients to testing would be beneficial, and one would therefore set θ to a large negative value, or to $-\infty$, corresponding to ASO. Conversely, if both ratios are small, resulting in a small β, then we expect that routing $*T$ patients to the physician would be beneficial, corresponding to a large θ value, or to $+\infty$, corresponding to NSO. In other words, we expect β to have a negative association with θ^*. In Section 5, we will see that θ^* can be reliably estimated as a linear function of $\ln\beta$.

4.2 Jackson Network Model

In order to illustrate possible unintended consequences of standing orders, we formulate an open Jackson network model (Chen and Yao 2001), with the same topology as the infinite-server model (see Figure 4). The difference is that now we assume a single dedicated server for each of Queues 1–3 and we assume that with probability $\zeta \in [0,1]$, standing orders are used for $*T$ patients—that is, $*T$ patients are routed to testing (Queue 2) after triage. Unlike the threshold policy (1), this randomized policy ignores congestion, but it permits a simple analysis of the network. Like the threshold policy, this randomized policy includes ASO (corresponding to $\zeta = 1$) and NSO (corresponding to $\zeta = 0$) as special cases.

To complete the specification of the Jackson network model, we assume a non-preemptive work-conserving service discipline at each node, i.i.d. exponentially-distributed service times with mean τ_i at Node $i, i = 1,2,3$, and an independent Poisson process with rate λ for arrivals to triage.

We begin by establishing a stability condition for the network.

Proposition 2. *The Jackson network model is stable if the following condition holds:*

$$\lambda < \min\left\{\frac{1}{\tau_1}, \frac{1}{\eta_{*T}\,\tau_2}, \frac{1}{\eta_{*T}\,\tau_3}\right\}. \tag{9}$$

Proof: Let u_i be the long-run average rate at which work arrives exogeneously to the system, destined for Node i. By tracing flows in Figure 4, we obtain:

$$u_1 = (\eta_{*O} + (1-\eta_{*O})\zeta)\,\lambda\tau_1,$$
$$u_i = \left(1 - \left(1 - \frac{\eta_{TT}}{\eta_{*T}}\right)\zeta\right)\eta_{*T}\lambda\tau_i,\ i = 2,3.$$

Note that η_{TT}/η_{*T} is the conditional probability that a patient who is triaged as target $(*T)$ is indeed a target patient. For Queue 1, (9) implies $\lambda\tau_1 < 1$. The quantity that pre-multiplies $\lambda\tau_1$ is the probability that a patient will have an initial evaluation and is therefore less than or equal to 1. It follows that $u_1 < 1$. Similarly, for Queues 2 and 3, (9) implies $\eta_{*T}\lambda\tau_i < 1$. The quantity that pre-multiplies $\eta_{*T}\lambda\tau_i$ is the probability that a patient will go through testing, and is therefore less than or equal to 1. Hence, $u_i < 1, i = 2,3$. It follows from Theorem 1.1 in Chang et al. (1994) that the network is stable. ■

Proposition 2 implies that under (9), the network is stable under ASO, NSO, and for any $\zeta \in (0,1)$. Proposition 2 holds even if the processing times and the inter-arrival times for new patients have general (as opposed to exponential) distributions. Permitted queue disciplines include priority disciplines, which one would be likely to see in an ED.

In the remainder of this subsection, we demonstrate numerically that an increase in LOS for the target patients need not result in a corresponding increase in the overall LOS and that the use of ASO may, surprisingly, lead to a decrease in LOS for over-tested patients. The average system time in Queue i is given by the $M/M/1$ formula

$$W_i^{\gamma,J} = \frac{\tau_i}{1-u_i}. \tag{10}$$

Substituting (10) into (4)-(6) provides values for Δ_p^J for $p \in \{OT, TT, OO\}$.

For the numerical experiments, we fixed the arrival rate ($\lambda = 1$), the probability that a patient is in the target population ($\eta_{TT} = \psi = 0.18$), the probability that a non-target patient is correctly identified at triage ($\nu = 0.83$), and the mean post-test evaluation time ($\tau_3 = 1$). We varied the mean initial evaluation time (τ_1) in the range $[0.2, 1]$ and we varied the mean testing time (τ_2) in the range $[0.2, 3]$. In Figures 6–7, we contrast predictions from the infinite-server and the Jackson network models about whether ASO or NSO will perform better. Specifically, in Figure 6, we show where the difference between the ASO and NSO overall LOS values, Δ_{**}^{∞}, changes sign, under the infinite-server model, whereas in Figure 7, we show where Δ_p^J changes sign, for various patient types p, under the Jackson network model. Recall that $\Delta_p^M > 0$ implies that model M predicts that NSO results in a lower LOS for patient type p. We observe the following from Figures 6–7:

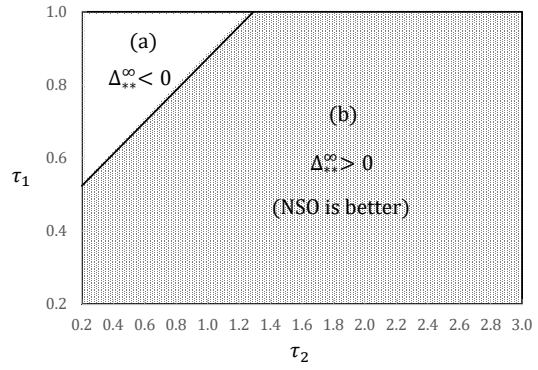

Figure 6 **Predictions for whether ASO or NSO results in lower LOS, for the $M = \infty$ model.**

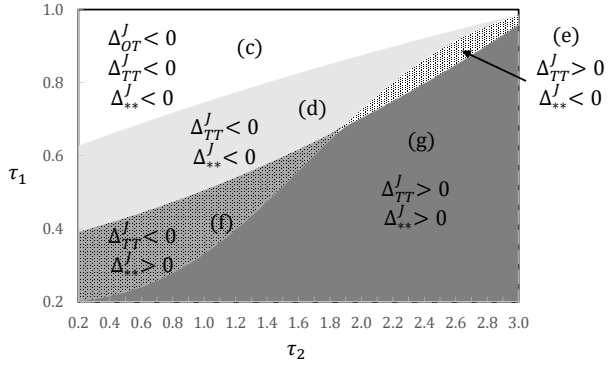

Figure 7 **Predictions for whether ASO or NSO results in lower LOS, for the $M = J$ model.**

Region (c): The LOS for overtested patients can *decrease* under ASO, if the initial exam time is long and the testing time is short.

Region (e): The LOS for target patients can increase under ASO, even if the overall LOS decreases, if the initial exam time is medium-to-long and the testing time is long.

Region (f): The overall LOS can increase under ASO, even if the LOS for target patients decreases, if both the initial exam time and the testing time are short.

Region (a) vs. Regions (c), (d), and (e): Comparison of Figures 6 and 7 shows that for these parameter values, the Jackson network model predicts a larger region where ASO is preferable in terms of overall LOS (the union of Regions (c), (d), and (e)) than the infinite-server model (Region (a)).

4.3 MDP Model

We formulate an MDP model to investigate the performance of the threshold policy (1), relative to an optimal policy. We assume the network topology in Figure 5, with a single physician serving Queues 1, 3, and 4 (thus, $Q_{\text{phys}} = Q_1 + Q_3 + Q_4$), and a single-server testing resource serving Queue 2 (thus, $Q_{\text{test}} = Q_2$). We let $Q_i(t)$ be the number of patients waiting or being served in Queue i at time t, with t often omitted, for brevity. Service times are independent and exponentially distributed, with rate $\mu_i = 1/\tau_i$, for Queue i, $i = 1, \ldots, 4$.

The model is Markovian. Full state information is assumed available at the time of triage routing: the number of patients in each queue, and the queue (if any) that the physician is serving. The state vector is $\mathbf{X} = \big(\mathbf{Q}, R\big) = ((Q_1, \ldots, Q_4), R)$, with $R = i, i \in \{1, 3, 4\}$ if the physician is serving Queue i and $R = 0$ if the physician is idle. The physician is idle if and only if $Q_{\text{phys}} = 0$ and the physician can serve Queue i only if $Q_i > 0, i \in \{1, 3, 4\}$.

These conditions define the set $\Omega(\mathbf{Q})$ of possible values for R given $\mathbf{Q}$. To focus attention on triage routing, we assume that the physician selects the next queue to serve randomly, selecting Queue i with probability Q_i/Q_{phys}, if $Q_{\text{phys}} > 0$.

A policy γ specifies whether to route $*T$ patients to Q_1 ("physician") or to Q_2 ("test"). We formulate the problem as an infinite-horizon average-cost MDP with a countably-infinite state space. The state vector under policy γ is $\mathbf{X}^\gamma = (\mathbf{Q}^\gamma, R^\gamma)$. The state space is $\mathcal{X} = \left\{ \mathbf{X} = (\mathbf{Q}, R) \,\middle|\, \mathbf{Q} \in \mathbb{Z}_+^4, R \in \Omega(\mathbf{Q}) \right\}$, with $\mathbb{Z}_+$ the set of non-negative integers. We seek a policy that minimizes the expected average number of patients in the system:

$$g^\gamma = L = \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}\Big[\int_0^T (Q_1^\gamma(t) + \cdots + Q_4^\gamma(t)) dt \Big]. \tag{11}$$

Under our assumptions, for any γ, $\mathbf{X}^\gamma$ is a uniformizable continuous-time Markov chain. We set $\Lambda = \lambda + \mu_1 + \cdots + \mu_4$, $\hat{\lambda} = \lambda/\Lambda$, and $\hat{\mu}_i = \mu_i/\Lambda$, $i = 1, \ldots, 4$, and let $\hat{\mathbf{X}}^\gamma(k)$ be the uniformized discrete-time Markov chain (with k often omitted, for brevity). Finding γ that minimizes (11) is equivalent to finding γ that minimizes

$$\hat{g}^\gamma = \frac{g^\gamma}{\Lambda} = \limsup_{K\to\infty} \frac{1}{K} \mathbb{E}\Big[\sum_{k=1}^{K} (Q_1^\gamma(k) + \cdots + Q_4^\gamma(k)) \Big]. \tag{12}$$

The average cost of the optimal policy, γ^*, is $\hat{g}^* = \min_\gamma \hat{g}^\gamma$. The optimal average cost and the bias function $w(\mathbf{Q}, R)$ satisfy the following equation:

$$\begin{aligned} w(\mathbf{Q}, R) = {} & \frac{Q_1 + \cdots + Q_4}{\Lambda} - \hat{g}^* \\ & + \eta_{*T} \cdot \hat{\lambda} \cdot \min\Big\{ w\big(\mathbf{Q} + \mathbf{q}^1, r(R,1)\big), w(\mathbf{Q} + \mathbf{q}^2, R) \Big\} \\ & + \eta_{*O} \cdot \hat{\lambda} \cdot w\big(\mathbf{Q} + \mathbf{q}^4, r(R,4)\big) \\ & + \hat{\mu}_2 \cdot \Big(w\big(\mathbf{Q} - \mathbf{q}^2 + \mathbf{q}^3, r(R,3)\big) \cdot \mathbb{1}(Q_2 > 0) + w(\mathbf{Q}, R) \cdot \mathbb{1}(Q_2 = 0) \Big) \\ & + \frac{\eta_{TT}}{\eta_{*T}} \cdot \hat{\mu}_1 \cdot \Big(c(\mathbf{Q} - \mathbf{q}^1 + \mathbf{q}^2) \cdot \mathbb{1}(R = 1) + w(\mathbf{Q}, R) \cdot \mathbb{1}(R \neq 1) \Big) \\ & + \frac{\eta_{OT}}{\eta_{*T}} \cdot \hat{\mu}_1 \cdot \Big(c(\mathbf{Q} - \mathbf{q}^1) \cdot \mathbb{1}(R = 1) + w(\mathbf{Q}, R) \cdot \mathbb{1}(R \neq 1) \Big) \\ & + \hat{\mu}_3 \cdot \big(c(\mathbf{Q} - \mathbf{q}^3) \cdot \mathbb{1}(R = 3) + w(\mathbf{Q}, R) \cdot \mathbb{1}(R \neq 3) \big) \\ & + \hat{\mu}_4 \cdot \big(c(\mathbf{Q} - \mathbf{q}^4) \cdot \mathbb{1}(R = 4) + w(\mathbf{Q}, R) \cdot \mathbb{1}(R \neq 4) \big), \qquad \forall\, (\mathbf{Q}, R) \in \mathcal{X}, \end{aligned} \tag{13}$$

where $\mathbf{Q} + \mathbf{q}^i$ corresponds to "add one patient to Queue i" and $\mathbf{Q} - \mathbf{q}^i$ corresponds to "remove one patient from Queue i." The function $r(R, i) = R \cdot \mathbb{1}(R \neq 0) + i \cdot \mathbb{1}(R = 0)$ models the assumption that the physician will begin serving a patient newly arrived to Queue i only if the physician is currently idle. The function $c(\mathbf{Q})$ models the physician's probabilistic choice of the next patient to serve, and is defined as:

$$c(\mathbf{Q}) = \begin{cases} \sum_{i \in \{1,3,4\}} \frac{Q_i}{Q_{\text{phys}}} w(\mathbf{Q}, i), & Q_{\text{phys}} > 0, \\ w(\mathbf{Q}, 0), & Q_{\text{phys}} = 0. \end{cases}$$

The meaning of the components of the right side of (13) is as follows. The first term is the difference between the number of patients in the system and the optimal average cost. The second term represents triage routing for $*T$ patients, with $w(\mathbf{Q}+\mathbf{q}^2, R)$ representing "test" and $w\big(\mathbf{Q}+\mathbf{q}^1, r(R,1)\big)$ representing "physician". The third term represents the arrival of a $*O$ patient to Queue 4. The fourth term represents a patient who completes testing and transfers to Queue 3. This can only happen if $Q_2 > 0$ and therefore, for uniformization, we add a self-transition term $w(\mathbf{Q}, R) \cdot \mathbb{1}(Q_2 = 0)$. The last four terms represent physician service completion. A patient receiving service in Queue 1 is a target patient with probability η_{TT}/η_{*T} and such a patient moves to Queue 2 after service completion. All other patients exit the system after service completion. If $R = i, i \in \{0,1,3,4\}$, then the physician completes service for Queue i with probability $\mathbb{1}(R \neq 0)\hat{\mu}_i$, and a self-transition occurs with probability $\sum_{j\in\{1,3,4\}, j\neq i} \hat{\mu}_j$.

We summarize properties of the MDP model in the following proposition.

PROPOSITION 3. *If* $\lambda < \min\left\{\frac{1}{\tau_1+\eta_{T*}\tau_3}, \frac{1}{\eta_{T*}\tau_2}\right\}$, *then there exists an average-cost optimal stationary policy for the MDP defined by (13) with an average cost that is independent of the initial state.*

Proof: See Appendix B.

The condition in Proposition 3 ensures that the system is stable under NSO. The utilizations of the physician and testing servers under this policy are $\frac{\lambda}{\tau_1+\eta_{T*}\tau_3}$ and $\frac{\lambda}{\eta_{T*}\tau_2}$, respectively. Therefore, the condition guarantees that under NSO the utilizations of both servers remain below one.

It is natural to conjecture that the optimal decision is *monotone* with respect to Q_{test} and Q_{phys}, meaning that:

- if "test" is optimal for $(Q_{\text{test}}, Q_{\text{phys}})$, then it is also optimal for $(Q_{\text{test}}, Q_{\text{phys}}+1)$ and $(Q_{\text{test}}-1, Q_{\text{phys}})$, and
- if "physician" is optimal for $(Q_{\text{test}}, Q_{\text{phys}})$, then it is also optimal for $(Q_{\text{test}}+1, Q_{\text{phys}})$ and $(Q_{\text{test}}, Q_{\text{phys}}-1)$.

Figures 8–9, which display the probability that "test" is optimal as a function of Q_{test} and Q_{phys} for two problem instances, shows that this is typically true. (We truncated the state space so that $Q_1+\cdots+Q_4 \leq 15$ and found that under the optimal policy, $\Pr\{Q_1+\cdots+Q_4 = 15\}$ was equal to 0.007 and 0.038 for the two instances, suggesting that the truncation had little effect on the results. Note that in the truncated model, no action is available at the boundary, where the system is full.)

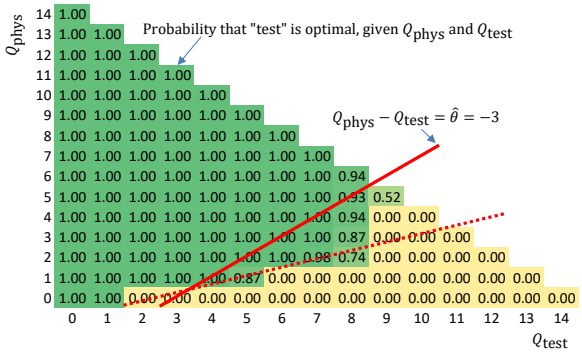

Figure 8 **MDP instance with $\lambda = 2.59, \tau_1 = \tau_3 = \tau_4 = 0.29, \tau_2 = 0.77, \eta_{TT} = 0.25$, and $\eta_{OT} = 0.05$; resulting in $\hat{\theta} = -3$ and a 0.3% optimality gap.**

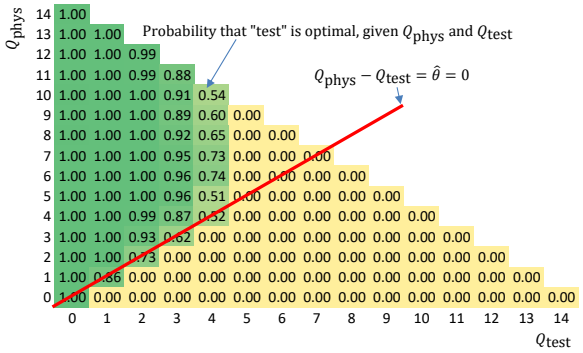

Figure 9 **Same MDP instance as in Figure 8, except $\tau_2 = 1.24$; resulting in $\hat{\theta} = 0$ and a 0.8% optimality gap.**

However, Figures 8–9 show that the optimal policy is complicated, in ways that could make it difficult to use in practice. First, the optimal action depends not only on $(Q_{\text{test}}, Q_{\text{phys}})$ but also on Q_1, Q_3, Q_4, and R, as evidenced by the fact that the probability that "test" is optimal is between 0 and 1 for some cells. (The state variables Q_1, Q_3, Q_4, and R vary within each cell, subject to the constraint $Q_{\text{phys}} = Q_1 + Q_3 + Q_4$.) Thus, the optimal policy consists of a 5-dimensional lookup table. Second, suppose that one were to approximate the optimal policy with a straight line in the space of Q_{test} and Q_{phys}, such as the dotted line $Q_{\text{phys}} = -(7/6) + (4/9)Q_{\text{test}}$ in Figure 8. This equation can be translated into a form similar to (1), namely $Q_{\text{phys}} - Q_{\text{test}} = -(7/6) - (5/9)Q_{\text{test}}$, but the right side of this latter equation is a *state-dependent threshold*, which complicates the use of the policy.

Fortunately, use of constant-threshold policies (represented by solid lines with a slope of +1; see Section 5 for how $\hat{\theta}$ is estimated), for the two instances illustrated in Figures 8–9, results in an overall LOS that is only 0.3% or 0.8% higher than optimal—despite these policies resulting in suboptimal actions for several cells. Section 5 shows that threshold policies of the form (1) are indeed near-optimal for most problem instances.

4.4 DES Model

The DES model uses the network topology in Figure 5. This model allows multiple servers for each queue, uses distributions that are fit to empirical data for processing times, and has a nonhomogenous Poisson arrival process with a time-dependent arrival rate obtained from real data. Patients waiting to see a physician (Queues 1 and 3) are seen in first-come-first-served order. The physicians are pooled, in the sense that a patient is seen by the first available physician, regardless of whether the patient has seen that physician previously.

Appendix F provides more detailed information about the DES model and Section 6 reports the results of our DES experiments.

5. MDP Experiments

In this section, we use the MDP model to numerically assess the optimality gap for the NSO, ASO, optimal-threshold ($\theta^* = \arg\min_\theta L(\theta)$), and approximate-threshold (with $\hat{\theta}$ as a linear function of $\ln\beta$) policies. The optimality gap is the percent increase in L relative to the optimal policy.

We find that the parameter β is predictive of the performance of threshold policies. In particular, we find that NSO and ASO perform poorly if $|\ln\beta| < 1$. In contrast, we find that the approximate threshold policy is nearly optimal.

5.1 Problem Instances

We generate a full factorial experiment of MDP problem instances by varying six factors: the average time the physician spends with a patient (τ), the target population proportion ($\eta_{TT} = \psi$), the initial examination duration as a proportion of the total time the physician spends with a target patient (κ), the overtesting rate under ASO (η_{OT}), and the utilization of resources under NSO ($u_{\text{phys}}^{\text{NSO}}$ and $u_{\text{test}}^{\text{NSO}}$). Table 4 lists the factor values. With the exception of τ, these factors are dimensionless. We fix the value of τ, without loss of generality. The factors τ, $\eta_{TT} = \psi$, κ, and η_{OT} correspond directly to quantities that are typically reported in medical studies, which helps us choose realistic values for those factors.

The experimental factors $\kappa, \tau, u_{\text{phys}}^{\text{NSO}}$, and $u_{\text{test}}^{\text{NSO}}$ are related to the model primitives via:

$$\tau = \tau_1 + \psi\tau_3, \qquad \kappa = \frac{\tau_1}{\tau_1 + \tau_3}, \qquad u_{\text{phys}}^{\text{NSO}} = \lambda\tau, \qquad u_{\text{test}}^{\text{NSO}} = \lambda\psi\tau_2.$$

Primitives that are not given directly can be computed from the experimental factors:

$$\lambda = \frac{u_{\text{phys}}^{\text{NSO}}}{\tau}, \qquad \nu = 1 - \frac{\eta_{OT}}{1-\psi},$$
$$\tau_1 = \frac{\kappa\,\tau}{\kappa + (1-\kappa)\psi}, \qquad \tau_2 = \frac{u_{\text{test}}^{\text{NSO}}}{u_{\text{phys}}^{\text{NSO}}}\frac{\tau}{\psi}, \qquad \tau_3 = \frac{(1-\kappa)\tau}{\kappa + (1-\kappa)\psi}.$$

The values for the experimental factors in Table 4 result in $3^5 = 243$ problem instances, which constitutes our *training set*. Table 12 in Appendix D shows minimum and maximum values for the model primitives and the resource utilizations under ASO ($u_{\text{phys}}^{\text{ASO}} = \lambda(\eta_{OO}\tau_1 + \eta_{*T}\tau_3)$ and $u_{\text{test}}^{\text{ASO}} = \lambda\eta_{*T}\tau_2$) across the instances in the training set. In Section 5.2, we use the training set to evaluate the performance of the optimal threshold policy and in Section 5.3 we use it to approximate the best threshold as a function of $\ln\beta$.

Parameter	Minimum	Baseline	Maximum	Reference(s)
Average time physician spends with a patient (τ)	—	22 min	—	Chonde et al. (2013)
Target population proportion ($\eta_{TT} = \psi$)	10%	25%	70%	Valtchinov et al. (2019), Compeau et al. (2016), Ghanes et al. (2015), Liu et al. (2014)
Initial examination duration as proportion of total service time for target patients (κ)	20%	35%	50%	Yang et al. (2016), Ellis et al. (2006), Graff et al. (1993)
Overtesting rate under ASO (η_{OT})	1%	5%	10%	Yang et al. (2016), Thurston and Field (1996), Lee et al. (1996), Davies (1994), Macleod and Freeland (1992)
Physician utilization under NSO ($u_{\text{phys}}^{\text{NSO}}$)	50%	80%	95%	Yang et al. (2016), Ellis et al. (2006), Graff et al. (1993)
Testing utilization under NSO ($u_{\text{test}}^{\text{NSO}}$)	50%	80%	95%	Steindel and Howanitz (2001), Edelstein et al. (2010)

Table 4 Experimental factors and their values.

To guard against overfitting, and in the spirit of the approaches used in Ehrhardt (1979) and Bravo and Shaposhnik (2020), we measure the performance of different estimation methods using a separate *test set*. The test set consists of additional problem instances generated by randomly selecting values for the experimental factors from uniform distributions with the minimum and maximum values in Table 4. We use the test set in Sections 5.3 and 5.4 to compare the performance of different policies.

As discussed in Appendix C, we truncate the state space by assuming that $Q_1 + \cdots + Q_4 \leq B = 15$. All instances are stable because of the truncation, even instances for which $u_{\text{testing}}^{\text{ASO}} > 1$ or $u_{\text{phys}}^{\text{ASO}} > 1$. The probability that the system is full, under the optimal policy, is 0.033 on average for the training set.

5.2 Optimal Threshold Policy

The optimality gap in the training set is only 0.5% on average and 2.4% in the worst case for the optimal threshold policy. Determining θ^* requires determining steady-state probabilities for $2B + 1$ Markov chains, corresponding to $\theta = -B, \ldots, +B$, with the size of each chain growing exponentially with B. It would therefore be valuable to avoid this computational effort by estimating θ^* directly from the model primitives. Figure 10 shows that, consistent with our discussion in Section 4.1, θ^* and $\ln \beta$ have a negative association. In the next section, we formulate and evaluate approximate threshold policies, in which θ^* is expressed as a linear or piece-wise linear function of $\ln \beta$.

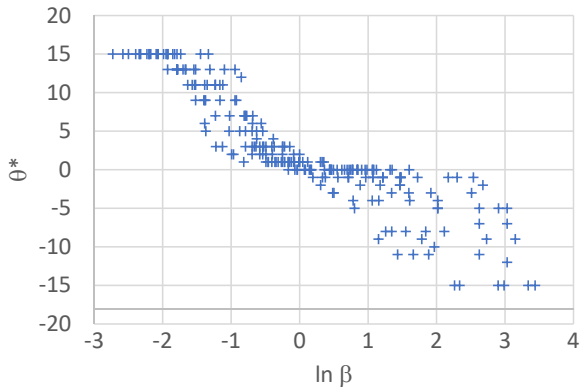

Figure 10 **Optimal threshold (θ^*) and $\ln\beta$ values in the training set.**

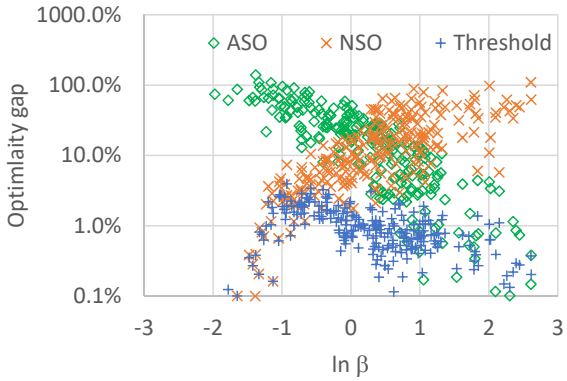

Figure 11 **Test-set optimality gap for NSO, ASO, and the approximate threshold policy.**

5.3 Approximate Threshold Policy

We use the $\ln\beta$ and θ^* values in the training set to estimate four regressions models. Table 13 in Appendix D shows complete results. For Models A and B, we utilize all 243 instances. Model A is linear and Model B is piecewise linear, obtained by estimating one regression models for the 133 instances with $\beta \leq 1$ and another for the 113 instances with $\beta \geq 1$. For Models C and D, we discard 33 instances for which $\theta^* = -B$ or $+B$, to reduce the impact of state-space truncation. Similar to Model B, the coefficients in Model D are estimated separately for 106 instances with $\beta \leq 1$ and for 107 instances with $\beta \geq 1$.

We compare the approximate threshold policies corresponding to the four regression models in terms of their test-set optimality gaps, as shown in the bottom half of Table 13. In computing the optimality gaps, threshold estimates are rounded to the nearest integer. Model D performs best, with a 1.1% average test-set gap (compared to 0.7% for the optimal threshold policy), and we use it for the remainder of our analysis:

$$\text{Model D:} \quad \hat{\theta} = \begin{cases} -0.15 - 7.22\ln\beta, & \beta \leq 1, \\ 0.90 - 2.79\ln\beta, & \beta > 1. \end{cases} \tag{14}$$

5.4 Extreme Routing Policies: NSO and ASO

The NSO policy routes all $*T$ patients to a physician. The ASO policy routes all $*T$ patients to testing. The medical literature that we reviewed indicates that these two extreme policies are the ones most frequently used in clinical settings.

Figure 11 shows the test-set optimality gaps for NSO, ASO, and the approximate threshold policy, as a function of $\ln\beta$. ASO performs well for $\ln\beta > 1$ and NSO performs well for $\ln\beta < -1$. Outside these ranges, the performance of ASO and NSO is abysmal, with optimality gaps exceeding 100% for many instances. The Model D approximate threshold policy, in contrast, has optimality gaps below 3.9%. This indicates that hospitals can

improve performance by using standing orders selectively, based on the number of patients waiting for testing and the number of patients waiting for a physician.

6. DES Experiments

In earlier sections, using a series of analytical models, we developed tentative insights regarding an effective threshold policy for routing $*T$ patients, how to estimate the threshold, and possible unintended consequences of using standing orders. These insights are contingent, however, on the strong assumptions we used to formulate tractable analytical models. In this section, we test the extent to which these insights hold in a more realistic setting, namely, in a DES model with a time-varying arrival rate, a time-varying number of physicians, and service time distributions that are based on real ED data.

We have one week of ED data from a mid-sized US hospital, consisting of records for 542 patient visits to the ED. Standing orders are not used in this ED, and therefore the data corresponds to a system using the NSO policy. A CT scan was ordered for 21% of the patients, and in our DES experiments we view this subgroup as the target population for a standing orders protocol. We discuss the data in detail in Appendix E. Here, we briefly outline the key model features that we have estimated from data.

Figure 12 shows the average number of patient arrivals and the scheduled number of physicians for each hour of the day. On average, 65.9 patients arrived per day and 2.3 physicians were on duty. We assume a non-homogeneous Poisson arrival process with the piece-wise constant rates shown in Figure 12. We use gamma distributions for the durations of the initial physician evaluation, testing, and post-test physician evaluation. Table 5 lists the means and squared coefficients of variation (SCVs) of the fitted distributions (fitting the distributions required the imputation of missing values; see Appendix E). We route $*T$ patients according to (1), with the threshold calculated as follows. First, we compute β using (7). Second, we use Model D (see (14)) to compute $\hat{\theta}$ and round the resulting value to the closest integer. For the base case, the calculations are as follows, assuming $\eta_{OT} = 0.01$:

$$\begin{aligned} \beta &= \left(1 + \frac{\eta_{TT}}{\eta_{OT}}\right) \frac{\tau_1}{\tau_2 + \tau_3} = \left(1 + \frac{0.212}{0.01}\right) \frac{26.24}{68.4 + 39.5} = 5.4 \\ \ln \beta &= 1.7 \\ \hat{\theta} &= 0.90 - 2.79 \ln \beta = -3.8, \text{ rounded to } \hat{\theta} = -4. \end{aligned}$$

These calculations do not require information about the arrival rate or the number of servers and therefore, the threshold does not vary with time.

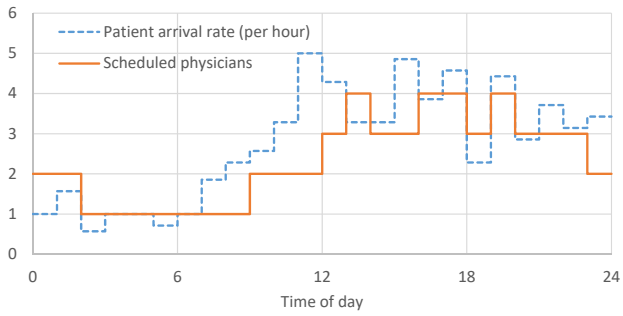

Figure 12 **Arrival rates and number of physicians on duty.**

Service duration (min.)	Mean	SCV
Phys. initial exam	26.2	0.754
Phys. post-test exam	39.5	0.754
CT scan test	68.4	0.423

Table 5 **Summary statistics for fitted gamma service time distributions.**

We use DES experiments to investigate how the impact of using standing orders depends on such factors as the overtesting rate, service time variability, on whether a threshold policy is used, the accuracy with which the threshold is estimated, and the level of ED congestion. All results are based on 250 thirty-day replications. The confidence interval half-width for the overall ED LOS for the base case, under NSO, is about 1.2%. For the case where we increase the service time variability the half-width increases to 2.3%.

In the base case, we vary the overtesting rate (η_{OT}): 1%, 5%, and 10% and we compare three policies: NSO, the approximate threshold policy, and ASO. Table 6 shows the results. We report the utilization of resources, the LOS for all patient types, and the improvement in overall LOS over NSO resulting from the use of the ASO and the threshold policy.

Note that under NSO, there are no OT patients. Under ASO, TT and OT patients should be treated the same. The fact that the average LOS is similar for these two patient types serves as an accuracy check for the DES model.

If we focus on overall ($**$) LOS, then we observe the following pattern: Using standing orders selectively, based on a threshold policy, reduces $**$ LOS for all levels of overtesting, but using standing orders for all patients reduces $**$ LOS if the rate of overtesting is low (1%) but not if it is medium and high (5% and 10%). In particular, if the triage nurse identifies 10% of patients as belonging to the target group even though they do not, then using standing orders for all patients increases the testing utilization from 66.0% to 95.7% and this results in more than doubling of the overall LOS (from 135.54 to 360.84 min.). In contrast, routing $*T$ patients to testing only when the testing resource is less congested, by using the threshold policy, reduces overall LOS, despite the high overtesting rate.

To check the robustness of the overall pattern of reduction in $**$ LOS, we investigate a series of variations of the base case. In each variation, we modify a small number of inputs but keep all other inputs the same.

Overtesting rate (η_{OT})	0%	1%		5%		10%	
Policy	NSO	ASO	Threshold	ASO	Threshold	ASO	Threshold
Physician utilization	69.9%	60.0%	60.5%	62.0%	64.6%	64.1%	67.7%
Testing utilization	66.0%	69.1%	69.3%	81.8%	76.1%	95.7%	77.8%
LOS (min.) by type							
$**$	135.54	99.04	97.98	142.58	115.40	360.84	130.61
TT	355.22	269.23	267.35	391.10	311.06	1043.41	355.70
OT	–	266.33	229.80	392.56	160.11	1042.45	128.21
OO	76.93	50.74	50.40	54.31	56.41	57.87	62.05
Improvement in $**$ LOS over NSO	–	26.9%	27.7%	-5.2%	14.9%	-166.2%	3.6%
Estimated threshold ($\hat{\theta}$)	–	–	-4	–	0	–	2
Proportion of $*T$ routed to testing	0%	100%	94.0%	100%	63.3%	100%	37.3%

Table 6 Comparison of different policies for the base case under different overtesting rates.

We begin by varying the SCVs of the three service time distributions: We compare a no variability (SCV = 0) case and a high variability (SCV = 2) case to the base case, in which the SCVs are 0.423 or 0.754 (see Table 5). The mean service times remain unchanged. Although varying the SCVs between 0 and 2 has a large impact on the LOS values (results not shown), Table 7 demonstrates that the pattern of percent improvement in $**$ LOS over NSO remain largely the same, regardless of the service time variability. The main exception appears to be that for the high variability case (SCV = 2) and for high overtesting (10%), the improvement from using the threshold policy almost disappears—but remains positive.

Overtesting rate (η_{OT})		1%		5%		10%	
Policy		ASO	Threshold	ASO	Threshold	ASO	Threshold
Improvement in $**$ LOS over NSO	Base case	26.9%	27.7%	-5.2%	14.9%	-166.2%	3.6%
	SCV = 2	26.4%	26.3%	-10.5%	14.2%	-147.0%	0.2%
	SCV = 0	23.2%	23.5%	-7.4%	11.8%	-179.5%	4.0%
Proportion of $*T$ routed to testing	Base case	100%	94.0%	100%	63.3%	100%	37.3%
	SCV = 2	100%	87.6%	100%	62.4%	100%	42.9%
	SCV = 0	100%	96.4%	100%	62.9%	100%	30.7%

Table 7 Changing SCVs of service times for the base case.

Next, we investigate the consequence of inaccurate estimation of the threshold used to determine how to route $*T$ patients. Table 8 shows that over- or underestimating the threshold $\hat{\theta}$ by one unit does not greatly impact the $**$ LOS. The pattern that the improvement from using the threshold policy decreases with the overtesting rate remains. We see

		Overtesting rate (η_{OT})		
	Threshold	1%	5%	10%
Improvement in ** LOS over NSO	$\hat{\theta}-1$	27.7%	14.4%	3.3%
	$\hat{\theta}$	27.7%	14.9%	3.6%
	$\hat{\theta}+1$	27.4%	12.6%	2.3%

Table 8 **Performance of the threshold policy for the correct threshold $\hat{\theta}$, and for the thresholds $\hat{\theta}\pm 1$, for the base case.**

		Overtesting rate (η_{OT})		
	Policy	1%	5%	10%
Improvement in ** LOS over NSO	Threshold	27.7%	14.9%	3.6%
	Random	26.4%	2.7%	-9.0%
	ASO	26.9%	-5.2%	-166.2%

Table 9 **Comparison of the threshold policy to a random routing policy.**

an asymmetry, in that overestimating $\hat{\theta}$ by one unit appears to be more harmful than underestimating by one unit, particularly for high overtesting rates.

Tables 6–8 show that as the overtesting rate increases, the threshold policy routes a smaller proportion of $*T$ patients to testing. One might wonder whether routing the "right" proportion to testing is sufficient to obtain the benefits of the threshold policy. To that end, we investigate the performance of a random routing policy, as in the Jackson network model, in which each $*T$ patient is routed to testing with probability ζ, independent of the system state. We set ζ equal to the proportion of $*T$ patient routed to testing by the threshold policy (see Table 6). Table 9 shows the results. We see that the random routing policy retains most of the performance benefit of the threshold policy if the overtesting rate is 1%, it retains some of the benefit if the overtesting rate is 5%, but it performs *worse* than NSO if the overtesting rate is 10%. This experiment demonstrates the importance of taking system congestion into account in routing policies—an aspect that is largely overlooked in the medical literature.

In our final set of experiments, we vary the average duration of initial evaluation and testing by ±25%. Table 10 shows the results, for a 5% overtesting rate. As before, use of the threshold policy reduces ** LOS both when mean service times are increased and when they are decreased, although the percent improvement is less than in the base case.

A closer look at the results in Table 10 reveals that if service times are decreased by 25% and ASO is used, target patients experience shorter LOS, but overall LOS increases. This outcome corresponds to Region (f) in Figure 7 and illustrates the importance of assessing the effect of the use of standing orders for the ED as a whole.

If service times are increased by 25% and the threshold policy is used, we see the opposite—target patients experience longer LOS, but overall, LOS decreases by 4.9%. This outcome corresponds to region (e) in Figure 7.

Initial evaluation and testing duration:	-25%			+25%		
Policy	NSO	ASO	Threshold	NSO	ASO	Threshold
Physician utilization	56.4%	51.2%	53.7%	82.2%	70.5%	76.4%
Testing utilization	49.6%	61.4%	55.5%	82.3%	97.6%	92.7%
LOS (min.) by type						
$**$	66.98	70.67	62.55	287.35	506.88	273.34
TT	190.44	183.69	170.85	724.68	1728.32	782.73
OT	–	184.38	77.53	–	1731.76	355.51
OO	34.01	30.63	30.62	170.35	95.45	122.80
Improvement in $**$ LOS over NSO	–	-5.5%	6.6%	–	-76.4%	4.9%
Estimated threshold ($\hat{\theta}$)	–	–	1	–	–	0
Proportion of $*T$ routed to testing	0%	100%	49.6%	0%	100%	39.9%

Table 10 Varying the mean initial evaluation and testing duration by $\pm 25\%$ with overtesting rates of 5%.

The medical studies that we reviewed in Section 2 appear to take the viewpoint that the aim is to reduce LOS and the means to do so is to reduce LOS for target patients. None of the studies report the overall LOS, however, or otherwise measure possible spillover effects on other patients. Our simulation results demonstrate that a negative finding for the target patients could, paradoxically, occur even if overall ED performance is improved.

Combining the results from Tables 6 and 10, for a 5% overtesting rate, we see that as we move from 25% lower service times to the base case to 25% higher service times, the testing utilization under NSO increases from 50% to 65% to almost 98% (physician utilization moves in the same direction, but not as drastically). The resulting percent improvements from using the threshold policy are 6.6%, 14.9%, and 4.9%. This suggests that if the system congestion is relatively low or high, there is less room for performance improvement via standing orders. In other words, the benefit from the optimal use of standing orders has an inverted U-shape relationship with resource utilization.

7. Conclusion

Standing orders allow an ED triage nurse to initiate certain medical tests for target patients before they are seen by a physician. In the medical literature, standing orders are viewed as a tool to reduce the ED LOS by reducing LOS for the target patients, without taking a system-wide viewpoint. Only a few studies report on how and if triage nurses ascertain the operational state of the ED in deciding whether to invoke standing orders.

We developed a series of models to investigate the impact of standing orders on the ED as a whole and derived a simple but near-optimal policy, which uses a single threshold

($\hat{\theta}$) to determine whether to use standing orders. The threshold policy recommends that patients triaged as target should be routed to testing if the difference between the number of patients waiting for a physician and the number of patients waiting for testing is greater than or equal to $\hat{\theta}$. We determine the value of $\hat{\theta}$ using a linear function of $\ln \beta$. The parameter β is a simple function of a subset of the model primitives, which excludes the arrival rate and the number of physicians on duty. Consequently, $\hat{\theta}$ does not depend on time of day or day of the week, which should simplify use of the policy. Numerical results for a simplified MDP model show that the performance of the approximate threshold policy is within 1.0% of an optimal policy.

The use of the approximate threshold policy requires information on the operational status of the ED, which could be obtained through direct observation of the queues in the ED, or could be available through a computerized information system. The continuing digitization of EDs facilitates availability of such information. For example, the Ministry of Health and Wellbeing of South Australia informs constituents virtually in real time of the operational status of the EDs in the region (Government of South Australia 2020). Their ED dashboard is updated every 30 minutes, providing expected ED waiting times, the number of patients waiting for consults, radiology services, inpatient beds, etc.

We used DES experiments to investigate in detail the impact of standing orders for different patient categories. We found that using standing orders could improve the performance, not by decreasing the ED LOS of the target patients, but rather through impacting other patients. This happens as routing some target patients to testing after triage decreases the physician load, which reduces wait times for physicians. We also found that the benefit of using the approximate threshold policy—rather than always or never using standing orders, regardless of operational status—is greater if resource utilizations are moderate.

The benefit of standing orders depends on how much their use can reduce the load on physicians without overwhelming testing resources. Our models focus on one standing order protocol. However, in practice, an ED could have multiple standing orders protocols. We predict that multiple standing order protocols, whose use is controlled with congestion-based policies will result in a larger reduction LOS reduction than any of the individual protocols: The increased load on testing will be distributed over multiple testing services, whereas the load on physicians is reduced even more. Investigating whether this is true is an important avenue for further research.

References

Ansari, S, SMR Iravani, Q Shao. 2019. Optimal control policies in service systems with limited information on the downstream stage. *Naval Research Logistics* **66**(5) 367–392.

Ashurst, JV, T Nappe, S Digiambattista, A Kambhampati, S Alam, M Ortiz, P Delpais, et al. 2014. Effect of triage-based use of the Ottawa foot and ankle rules on the number of orders for radiographic imaging. *The Journal of the American Osteopathic Association* **114**(12) 890–891.

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

Bertsimas, D., D. Gamarnik, J.N. Tsitsiklis. 1996. Stability conditions for multiclass fluid queueing networks. *IEEE Transactions on Automatic Control* **41**(11) 1618–1631.

Bramson, M. 2008. Stability of queueing networks. *Probability Surveys* **5** 169 – 345.

Bravo, F, Y Shaposhnik. 2020. Mining optimal policies: A pattern recognition approach to model analysis. *INFORMS Journal on Optimization* **2**(3) 145–166.

Castner, J, S Grinslade, J Guay, AZ Hettinger, JY Seo, L Boris. 2013. Registered nurse scope of practice and ed complaint-specific protocols. *Journal of Emergency Nursing* **39**(5) 467–473.

Chang, C, JA Thomas, S Kiang. 1994. On the stability of open networks: A unified approach by stochastic dominance. *Queueing Systems* **15**(1–4) 239–260.

Chen, H, DD Yao. 2001. Jackson networks. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer, New York, 15–35.

Cheung, WWH, L Heeney, JL Pound. 2002. An advance triage system. *Accident and Emergency Nursing* **10**(1) 10–16.

Chonde, S, C Parra, C Chang. 2013. Minimizing flow-time and time-to-first-treatment in an emergency department through simulation. *Proceedings of the 2013 Winter Simulation Conference*. IEEE, 2374–2385.

Cochran, JK, KT Roche. 2009. A multi-class queuing network analysis methodology for improving hospital emergency department performance. *Computers & Operations Research* **36**(5) 1497–1512.

Compeau, S, M Howlett, S Matchett, J Shea, J Fraser, R McCloskey, P Atkinson. 2016. Does elimination of a laboratory sample clotting stage requirement reduce overall turnaround times for emergency department stat biochemical testing? *Cureus* **8**(10) e819.

Corl, K. 2019. Hospitals' new emergency department triage systems boost profits but compromise care. *Stat* `www.statnews.com/2019/09/05/triage-system-boost-profits-compromises-care/`, accessed 2020-05-25.

Davies, J. 1994. X-ray vision of shorter queues. *Nursing Times* **90**(21) 52.

Delasay, M, A Ingolfsson, B Kolfal. 2016. Modeling load and overwork effects in queueing systems with adaptive service rates. *Operations Research* **64**(4) 867–885.

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

Edelstein, WA, M Mahesh, JA Carrino. 2010. MRI: time is dose—and money and versatility. *Journal of the American College of Radiology* **7**(8) 650–652.

Ehrhardt, R. 1979. The power approximation for computing (s, S) inventory policies. *Management Science* **25**(8) 777–786.

Ellis, DG, J Mayrose, M Phelan. 2006. Consultation times in emergency telemedicine using realtime videoconferencing. *Journal of Telemedicine and Telecare* **12**(6) 303–305.

Fan, J, K Woolfrey. 2006. The effect of triage-applied Ottawa Ankle Rules on the length of stay in a Canadian urgent care department: A randomized controlled trial. *Academic Emergency Medicine* **13**(2) 153–157.

Ghanes, K, O Jouini, M Wargon, Z Jemai. 2015. Modeling and analysis of triage nurse ordering in emergency departments. *2015 International Conference on Industrial Engineering and Systems Management*. IEEE, 228–235.

Goldstein, L, M Wells, C Vincent-Lambert. 2018. A randomized controlled trial to assess the impact of upfront point-of-care testing on emergency department treatment time. *American Journal of Clinical Pathology* **150**(3) 224–234.

Government of South Australia, Ministry of Health and Wellbeing. 2020. Emergency department dashboard. `www.sahealth.sa.gov.au/wps/wcm/connect/public+content/sa+health+internet/about+us/our+performance/our+hospital+dashboards/about+the+ed+dashboard/emergency+department+dashboard`. Accessed 2020-05-25.

Graff, LG, S Wolf, R Dinwoodie, D Buono, D Mucci. 1993. Emergency physician workload: A time study. *Annals of Emergency Medicine* **22**(7) 1156–1163.

Harrison, JM, AJ Lemoine. 1981. A note on networks of infinite-server queues. *Journal of Applied Probability* **18**(2) 561–567.

Ho, JK, JP Chau, JT Chan, CH Yau. 2018. Nurse-initiated radiographic-test protocol for ankle injuries: A randomized controlled trial. *International Emergency Nursing* **41** 1–6.

Hu, X, S Barnes, B Golden. 2018. Applying queueing theory to the study of emergency department operations: a survey and a discussion of comparable simulation studies. *International Transactions in Operational Research* **25**(1) 7–49.

Huang, J, B Carmeli, A Mandelbaum. 2015. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research* **63**(4) 892–908.

Hwang, CW, T Payton, E Weeks, M Plourde. 2016. Implementing triage standing orders in the emergency department leads to reduced physician-to-disposition times. *Advances in Emergency Medicine* **2016**.

Kalra, MG, KE Higgins, ED Perez. 2016. Common questions about streptococcal pharyngitis. *American Family Physician* **94**(1) 24–31.

Kamali, MF, T Tezcan, O Yildiz. 2018. When to use provider triage in emergency departments. *Management Science* **65**(3) 1003–1019.

Lee, KM, TW Wong, R Chan, CC Lau, YK Fu, KH Fung. 1996. Accuracy and efficiency of X-ray requests initiated by triage nurses in an accident and emergency department. *Accident and Emergency Nursing* **4**(4) 179–181.

Lee, WW, L Filiatrault, RB Abu-Laban, A Rashidi, L Yau, N Liu. 2016. Effect of triage nurse initiated radiography using the Ottawa Ankle Rules on emergency department length of stay at a tertiary centre. *Canadian Journal of Emergency Medicine* **18**(2) 90–97.

Li, Y, Q Lu, H Du, J Zhang, L Zhang. 2018. The impact of triage nurse-ordered diagnostic studies on pediatric emergency department length of stay. *The Indian Journal of Pediatrics* **85**(10) 849–854.

Lindley-Jones, M, BJ Finlayson. 2000. Triage nurse requested X-rays—are they worthwhile? *Emergency Medicine Journal* **17**(2) 103–107.

Liu, S, B Liu, HB Xiao. 2014. The utilisation of ECG in the emergency department. *The British Journal of Cardiology* **21** 1–2.

Macleod, AJ, P Freeland. 1992. Should nurses be allowed to request X-rays in an accident & emergency department? *Emergency Medicine Journal* **9**(1) 19–22.

Nestler, DM, AR Fratzke, CJ Church, L Scanlan-Hanson, AT Sadosty, MP Halasy, J L Finley, et al. 2012. Effect of a physician assistant as triage liaison provider on patient throughput in an academic emergency department. *Academic Emergency Medicine* **19**(11) 1235–1241.

Parris, W, S McCarthy, AM Kelly, S Richardson. 1997. Do triage nurse-initiated X-rays for limb injuries reduce patient transit time? *Accident and Emergency Nursing* **5**(1) 14–15.

Pedersen, GB, JO Storm. 2009. Emergency department X-rays requested by physicians or nurses. *Ugeskrift for Laeger* **171**(21) 1747–1751.

Puterman, ML. 2014. *Markov Decision Processes.: Discrete Stochastic Dynamic Programming*. Wiley.

Retezar, R, E Bessman, R Ding, SL Zeger, ML McCarthy. 2011. The effect of triage diagnostic standing orders on emergency department treatment time. *Annals of Emergency Medicine* **57**(2) 89–99.

Rosmulder, RW, JJ Krabbendam, AH Kerkhoff, ER Schinkel, LF Beenen, JS Luitse. 2010. "Advanced triage" improves patient flow in the emergency department without affecting the quality of care. *Nederlands Tijdschrift Voor Geneeskunde* **154** A1109–A1109.

Rowe, BH, C Villa-Roel, X Guo, MJ Bullard, M Ospina, B Vandermeer, G Innes, et al. 2011. The role of triage nurse ordering on mitigating overcrowding in emergency departments: A systematic review. *Academic Emergency Medicine* **18**(12) 1349–1357.

Rui, P, K Kang. 2017. National Hospital Ambulatory Medical Care Survey: 2017 emergency department summary tables. `www.cdc.gov/nchs/data/nhamcs/web_tables/2017_ed_web_tables-508.pdf` Last accessed on 2020-05-07.

Russ, S, I Jones, D Aronsky, RS Dittus, CM Slovis. 2010. Placing physician orders at triage: The effect on length of stay. *Annals of Emergency Medicine* **56**(1) 27–33.

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

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

Settelmeyer, D. 2018. Evaluation of an evidence-based throat-pain protocol to reduce left-without-being-seen, length of stay, and antibiotic prescribing. *Journal of Emergency Nursing* **44**(3) 236–241.

Shell, IG, GH Greenberg, RD McKnight, RC Nair, I McDowell, M Reardon, JP Stewart, et al. 1993. Decision rules for the use of radiography in acute ankle injuries: Refinement and prospective validation. *Journal of American Medical Association* **269**(9) 1127–1132.

Sigman, K. 1990. The stability of open queueing networks. *Stochastic Processes and their Applications* **35** 11–25.

Steindel, SJ, PJ Howanitz. 2001. Physician satisfaction and emergency department laboratory test turnaround time: Observations based on College of American Pathologists Q-Probes studies. *Archives of Pathology & Laboratory Medicine* **125**(7) 863–871.

Than, KC, YL Leong, BS Ngiam. 1999. Initiation of X-rays by the triage nurse: Competency and its effect on patients' total time spent in the accident and emergency department. *Annals of Emergency Medicine* **34**(4) S60.

Thurston, J, S Field. 1996. Should accident and emergency nurses request radiographs? Results of a multi-centre evaluation. *Emergency Medicine Journal* **13**(2) 86–89.

Valtchinov, VI, IK Ip, R Khorasani, JD Schuur, D Zurakowski, J Lee, AS Raja. 2019. Use of imaging in the emergency department: Do individual physicians contribute to variation? *American Journal of Roentgenology* **213**(3) 637–643.

Wiler, JL, S Welch, J Pines, J Schuur, N Jouriles, S Stone-Griffith. 2015. Emergency department performance measures updates: Proceedings of the 2014 Emergency Department Benchmarking Alliance Consensus Summit. *Academic Emergency Medicine* **22**(5) 542–553.

Yang, KK, SSW Lam, JMW Low, MEH Ong. 2016. Managing emergency department crowding through improved triaging and resource allocation. *Operations Research for Health Care* **10** 13–22.

Zayas-Caban, G, J Xie, LV Green, ME Lewis. 2019. Policies for physician allocation to triage and treatment in emergency departments. *IISE Transactions on Healthcare Systems Engineering* **9**(4) 342–356.

Zègre-Hemsey, JK, JJ Garvey, MG Carey. 2016. Cardiac monitoring in the emergency department. *Critical Care Nursing Clinics* **28**(3) 331–345.

A. Summary of Empirical Medical Literature

Reference	Location	Duration (months)	LOS reduction for target population
Thurston and Field (1996)	UK	NR	Overall: 4 min (4.3%)[a], Sent for X-rays: 14 min (29.2%)[a]
Parris et al. (1997)	Australia	5.5	With a fracture: 14 min[a], Without a fracture: 6 min[a]
Lindley-Jones and Finlayson (2000)	UK	0.5	37.2 min (36%)*** in time from triage to treatment decision
Fan and Woolfrey (2006)	Canada	3	6.7 min (8.4%)[a]
Lee et al. (2016)	Canada	12	28 min (19.6%)**
Ho et al. (2018)	Hong Kong	NR	13 min (14.9%)*
Lee et al. (1996)	Hong Kong	3	Sent for X-rays: 18.59 min***
Pedersen and Storm (2009)	Denmark	NR	For 75% patients: 21 min (60%) in time from admittance to X-rays request; 24 min (26.6%) in time from admittance to patient returned from X-ray
Rosmulder et al. (2010)	Netherlands	0.75	Overall: 14 min (14%); Those who require additional diagnostic investigation: 27 min (18%)
Ashurst et al. (2014)	USA	10	6.5 min (6.3%)[a]
Than et al. (1999)	Singapore	3	24.5 min
Hwang et al. (2016)	USA	5	−212 min (−52.7%)***; 26 min (16.9%)* in time from physician evaluation to disposition time if all tests were completed before the patient was seen by a physician
Li et al. (2018)	China	5	15 min (6.2%)***
Settelmeyer (2018)	USA	3	6 min
Cheung et al. (2002)	Canada	NR	46 min
Retezar et al. (2011)	USA	32	52 min (18%) in time between placement of the patient in treatment room and disposition decision
Goldstein et al. (2018)	South Africa	5.5	For a subset of tests: 20%* in time from first physician evaluation to disposition decision

Table 11 **Additional information from the medical studies on the impact of standing orders initiated by triage nurse.**
Legend: NR = not reported.
Legend for statistical significance of LOS reduction: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, a: not statistically significant, no code: statistical significance not reported.

B. Proof of Proposition 3

We prove Proposition 3 using Theorem 8.10.7 for countable–state MDPs from Puterman (2014), by verifying that (a) the immediate reward, with the MDP formulated as a maximization problem, is bounded above, and (b) the assumptions of Theorem 8.10.9 in

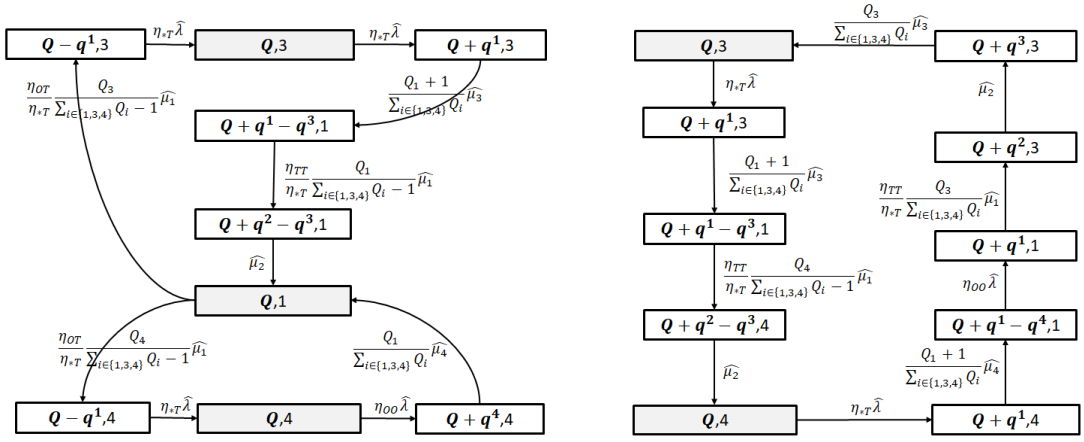

Figure 13 **Transition paths from $(\mathbf{Q}, R)$ to $(\mathbf{Q}, R')$, for all pairs $(R, R') \in \{1,3,4\}^2$**

Puterman (2014) hold.

Part (a): We redefine the MDP in (13) as a maximization problem with immediate reward defined as $r(s,a) = r(s) = r(\mathbf{Q}, R) = -\sum_i Q_i$. Because $-\sum_i Q_i \leq 0$, the assumption that $r(s,a)$ is bounded above is satisfied. In this proof, we work with policies that *maximize* the average reward, rather than minimize the average cost. Our immediate rewards are the negative of the total number of patients in the system. Therefore, the average reward is non-positive: if under some policy, the average number in the system is L, then the average reward is $-L$.

Part (b): Next, we verify the assumptions of Theorem 8.10.9. First, we show that there exists a deterministic stationary policy that induces a positive-recurrent Markov chain. To do so, we show that there exists a policy, namely, NSO, which results in a stable queueing system and a Markov chain in which all states communicate (Bramson 2008). To prove that all states of the NSO-induced Markov chain communicate, we show in Figure 13 that there is a non-zero probability of transitioning in a finite number of steps from any feasible origin state $(\mathbf{Q}, R)$ to any feasible destination state $(\mathbf{Q}, R')$ that differs only in the queue that the physician is serving (state variable R). Further, in Figure 14 we show that any feasible state $(\mathbf{Q} + \mathbf{q}^i, R')$ is reachable from $(\mathbf{Q}, 1)$ in a finite number of steps, that any state $(\mathbf{Q} - \mathbf{q}^R, R')$ is reachable from $(\mathbf{Q}, R)$ in a single step, and that any state $(\mathbf{Q} - \mathbf{q}^2, R')$ is reachable from $(\mathbf{Q}, 3)$ in two steps. We conclude that by combining such "atomic state changes" ($\mathbf{Q}$ to $\mathbf{Q} \pm \mathbf{q}^i$ and R to R'), any feasible state is reachable from any other feasible state, and therefore all states communicate.

Next, we note that the queueing network of interest is a two-station multi-class network in which the servers follow a work-conserving policy, and in which one of the stations

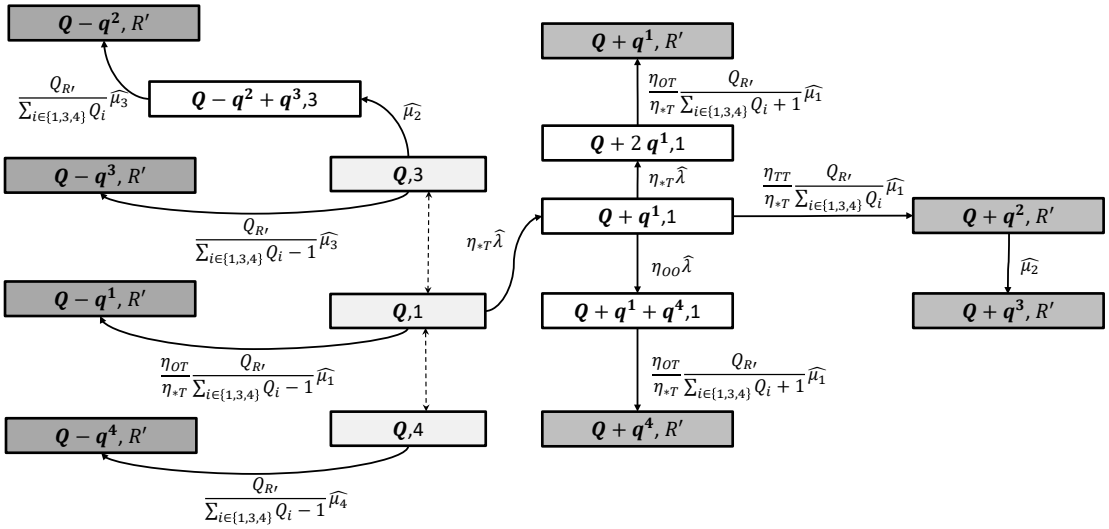

Figure 14 **Selected transition paths from $(\mathbf{Q}, R)$ to $(\mathbf{Q} - \mathbf{q}^R, R')$, from $(\mathbf{Q}, 3)$ to $(\mathbf{Q} - \mathbf{q}^2, R')$, and from $(\mathbf{Q}, 1)$ to $(\mathbf{Q} + \mathbf{q}^i, R')$. Details of transition paths between $(\mathbf{Q}, R)$ and $(\mathbf{Q}, R')$, indicated here with dashed lines, are shown in Figure 13.**

serves only one class. The two stations are the physician server and the testing server; the four classes are the patients in the four queues; and the station serving a single class is the testing station. Bertsimas et al. (1996, Theorem 4) proved that such a queueing network is stable if the total load on each station is less than 1. Under the NSO policy, this requirement translates into an upper limit on the total arrival rate:

$$\lambda < \min\left\{\frac{1}{\tau_1 + \eta_{T*}\,\tau_3}, \frac{1}{\eta_{T*}\,\tau_2}\right\}. \tag{15}$$

This concludes our verification of the first assumption of Theorem 8.10.9: provided that (15) is satisfied, there exists a deterministic stationary policy that induces a positive recurrent Markov chain.

Stability of a queueing system is defined in terms of finite expected queue lengths for all stations (Sigman 1990). System stability under NSO implies that g^{NSO} is finite, which is the second assumption of Theorem 8.10.9.

The last assumption of Theorem 8.10.9 is that the set $Y(g^{\text{NSO}}) = \{(\mathbf{Q}, R) \in \mathcal{X} : -\sum_i Q_i > g^{\text{NSO}}\}$ is non-empty and finite. To show that $Y(g^{\text{NSO}})$ is non-empty, we observe that $(\mathbf{0}, 0) \in Y(g^{\text{NSO}})$, because $g^{\text{NSO}} < 0$ (recall that g^{NSO} is an average of negative rewards). To show that $Y(g^{\text{NSO}})$ is finite, we observe that g^{NSO} is finite and that the Q_i are non-negative integers, and therefore the number of states that satisfy $\sum_i Q_i < -g^{\text{NSO}}$ is finite. ■

C. Calculation of Optimal Policy and Optimal Threshold Policy

For the optimal policy results reported in Section 5, we modified the MDP from Section 4 by truncating the state space to $Q_1 + \cdots + Q_4 \leq B = 15$, and replacing arrivals with self-transitions for states with $Q_1 + \cdots + Q_4 = B$. We solved the modified MDP using the relative value iteration algorithm, implemented using Python 3.6, with the stopping criterion $\epsilon = 10^{-5}$ and a maximum of 10,000 iterations.

We evaluated system performance under the threshold policy defined in (1) by constructing the transition probability matrix $P(\theta)$ for the Markov chain induced by a given threshold θ. We obtained the stationary probability vector π for this Markov chain by solving the system $A\pi = b$, using the SciPy Python library (Release 1.4.1), with the matrix A set to $P(\theta)^T - I$ and the last row replaced by a row of ones, and with b set to $(0, \ldots, 0, 1)^T$. We computed $L(\theta)$ as $\sum_{\mathbf{X}} \pi(\mathbf{X}) \left(Q_1(\mathbf{X}) + \cdots + Q_4(\mathbf{X})\right)$, where $\mathbf{X}$ indexes states. We obtained an optimal threshold θ^* as $\arg\min_{\theta \in \{-B, \ldots, +B\}} L(\theta)$.

D. Additional Results from MDP Experiments

Table 12 shows minimum and maximum values for the model primitives and the resource utilizations under ASO across all the problem instances in the training set used in the MDP experiments in Section 5.

Parameter	Minimum	Maximum
Arrival rate (λ)	1.36	2.59
Initial examination service time(τ_1)	5.8	20
Testing processing duration(τ_2)	16.5	428.6
Post-test examination service time (τ_3)	12.9	63.2
True negative rate (ν)	66.7%	98.9%
Physician utilization under ASO ($u_{\text{phys}}^{\text{ASO}}$)	29.4%	108.6%
Testing utilization under ASO ($u_{\text{test}}^{\text{ASO}}$)	50.7%	190.0%

Table 12 **Minimum and maximum values for the model primitives and resource utilizations under ASO, derived from experimental factor values. All service processing times are in minutes.**

The top half of Table 13 shows complete results for the four regression models we evaluate for estimating θ^* as a function of $\ln \beta$. We compare the approximate threshold policies corresponding to the four regression models in terms of their test-set optimality gaps, as shown in the bottom half of Table 13.

		Approximate threshold ($\hat{\theta}$)				Optimal
		Model A	Model B	Model C	Model D	threshold (θ^*)
$\beta \leq 1$	Intercept	2.70 (0.19)	0.10 (0.33)	2.39 (0.19)	-0.15 (0.36)	N/A
	Slope	-4.75 (0.13)	-7.01 (0.26)	-4.10 (0.16)	-7.22 (0.37)	
$\beta \geq 1$	Intercept	as above	1.51 (0.48)	as above	0.90 (0.43)	
	Slope		-3.62 (0.32)		-2.79 (0.31)	
$\beta \leq 1/\beta \geq 1$	R^2	0.84	0.85/0.54	0.76	0.78/0.44	
Test-set optimality gap	Minimum	0.0%	0.0%	0.0%	0.0%	0.0%
	Average	1.5%	1.1%	1.4%	1.1%	0.7%
	Median	1.3%	0.9%	1.2%	0.9%	0.6%
	Standard deviation	1.2%	0.7%	1.0%	0.7%	0.5%
	Maximum	7.2%	3.9%	5.3%	3.9%	2.2%
	Winning policy*	35.4%	68.7%	50.6%	70.0%	N/A

Table 13 Four models for estimation of θ^*. Standard errors are shown in parentheses.
*** The percentages do not add up to 100% because of ties.**

E. ED Data

We have records for 542 patients treated at a mid-sized US hospital over the course of one week. We removed 7 patient records because of missing data. Table 14 describes the variables in the data set.

We did not have access to physician schedules but we knew who the treating physician was for every patient. Using this information, we inferred the shift start and end time for each physician. We inferred the following shifts: 8-hour shifts starting at 6 am and at noon, a 9-hour shift starting at 9 am, 10-hour shifts starting at 1 pm and at 4 pm, and an 11-hour shift starting at 7 pm. Using these shifts, we calculated the scheduled number of doctors for each hour of day, as shown in Figure 12.

We require information about initial physician evaluation, post-test physician evaluation, and testing, but the data does not provide direct measurements of the durations of these activities. In the remainder of this section, we discuss how we imputed values for these durations.

Physician initial and post-test evaluations: Our procedure for imputing initial evaluation durations is based on the principle that during a busy period for a particular server, the server moves from one activity to the next with no delay. For each patient, we have some or all of the following timestamps:

AR: Patient arrives to the ED.

IE: Physician initial evaluation begins.

Description	Data type	Comments
Visit identifier	Numeric ID	542 patient visits
AR = Patient arrival	Timestamp	
Primary complaint	String	298 unique values; non-standardized free text, e.g. "RLQ ABDOMINAL PAIN, FEVER, HA"
IE = Physician evaluation	Timestamp	Assumed to be recorded at the beginning of initial evaluation
Evaluating physician	Physician ID	14 different physicians worked in the ED during the week for which the data was collected
DC = Discharge	Timestamp	
Scans ordered	Numeric	115 patients (21% of 542) had one or more CT scans ordered
CT scan type	String	22 unique values; The three most commonly ordered scan types represented 30%, 19%, and 13% of all scans ordered
OR = CT scan order	Timestamp	Assumed to be recorded when CT scan order is placed
RP = CT scan report completion	Timestamp	

Table 14 Variables in the ED data set.

Interval	Count	Min.	Avg.	Median	Max.	Std. Dev.	SCV	Time unit
Initial evaluation (imputed)	392	1	26.2	20	161	22.0	0.704	minutes
CT scan duration (imputed)	17	19	68.4	49	275	59.9	0.766	minutes

Frequency distribution for Scans ordered	
number of scans:number of patients	$\geq$ 1:115, 1:93, 2:16, 3:5, 4:1

Table 15 Summary statistics for the ED data.

OR: CT scan ordered (separate timestamp for each scan).

RP: CT scan report complete (separate timestamp for each scan report).

DC: Patient discharged.

We assume that IE, OR, and DC require the immediate attention of the patient's physician, but AR and RP do not. The imputation procedure is as follows:

1. For each Physician p and for each physician evaluation time stamp t_1, corresponding to Patient q: Identify the next timestamp t_2 of any type for Physician p. The interval (t_1, t_2) is a potential sample of an initial evaluation duration.
2. Eliminate interval (t_1, t_2) obtained in Step 1 if any of the following holds:
   (a) Timestamp t_2 is AR or RP, or timestamp t_2 is OR for Patient q. In words, we only keep intervals for which t_2 corresponds to physician evaluation or discharge (IE or DC) for any patient under Physician p's care, or ordering a test (OR) for any patient other than Patient q under Physician p's care.

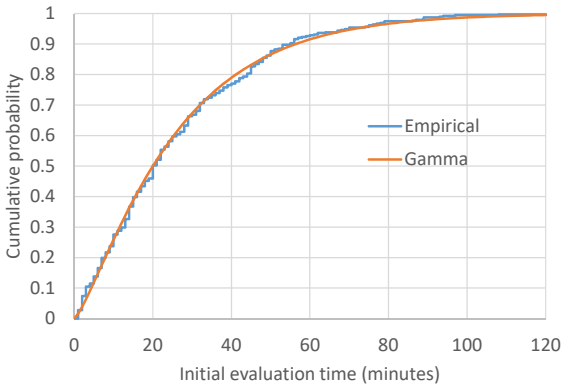

Figure 15 **Distribution of imputed initial evaluation durations.**

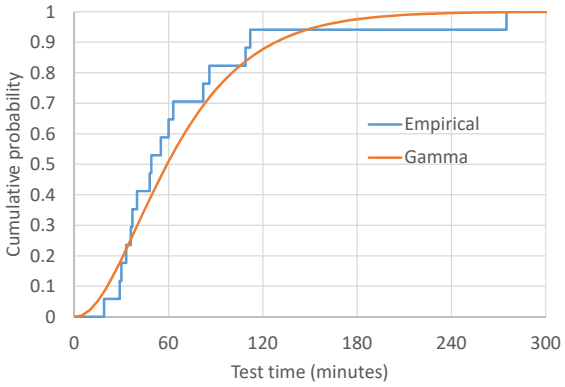

Figure 16 **Distribution of imputed test durations.**

(b) The duration of (t_1, t_2) is zero.

(c) The number of patients (not counting Patient q) that were waiting for a physician initial evaluation during (t_1, t_2) dropped below 1. In words, we eliminate intervals for which there is a possibility that Physician p was idle between completing one initial evaluation and starting their next task.

This procedure allows us to impute initial evaluation durations for 392 patients. The resulting summary statistics are provided in Table 15. The SCV of 0.704 is considerably less than 1—the value for an exponential distribution. Figure 15 shows that a gamma distribution provides a good fit to the empirical distribution. The maximum-likelihood parameter estimates for the gamma distribution are $\hat{k} = 1.325$ and $\hat{\theta} = 19.81$, resulting in a mean of $\hat{k}\hat{\theta} = 26.24$ minutes and SCV $= 1/\hat{k} = 0.754$ for the fitted distribution. We use this gamma distribution in the DES base case.

Our data has no timestamps that allow us to impute the durations of post-test evaluations for individual patients. We assume that post-test evaluations are 50% longer than initial evaluations, on average, and that the distribution shape is the same for post-test evaluations as initial evaluations. Thus, we assume a gamma distribution with parameters $k = 1.33$ and $\theta = 19.81 \times 1.5$, which leads to a mean of $k\theta = 39.5$ minutes and SCV $= 1/k = 0.754$. Assuming a proportion $\eta_{T*} = 115/542 = 21.2\%$ of target patients and NSO, the base case simulated physician utilization is 70%.

CT scan duration: A total of 144 CT scans were ordered, for 115 of the 542 patients. The number of scans per patient, for patients with one or more scans, ranged from 1 to 4 (see Table 15). Each CT scan had separate timestamps for the time of order and the time of report completion. We combined all scans for a single patient into a single "test",

with the test order time set to the earliest scan order time and the report completion time set to the latest scan report completion time. There were five exceptions, for which one or more of the scan order times for the patient occurred after one or more of the scan report completion times for the patient. For each of those five patients, we defined two or more tests, with each test consisting of one or two scans. This procedure resulted in 120 test duration intervals. The hospital had a single CT scanner and we assume that a single radiologist was on duty at all times, to prepare the scan reports. Therefore, each of these intervals corresponds to the sojourn time in a two-station tandem queueing system, where the first station is the single CT scanner and the second station is the single radiologist who prepares the CT scan reports.

Next, for each test duration interval, call it (t_1, t_2), we eliminated the interval if one or more other tests were completed within (t_1, t_2), which indicates that part of (t_1, t_2) could represent waiting time for the CT scanner or the radiologist. This reduced the number of tests from 120 to 45.

Finally, for each of the remaining test duration intervals, we kept only intervals for which at least one other test was ordered before t_1 and completed after t_2, indicating that the test of interest was given priority. This reduced the number of tests from 45 to 17. We assume that these 17 tests were completed with minimal delay, and therefore their durations are unbiased estimates of the total processing time for performing the CT scans and preparing the reports, in the aforementioned tandem queueing system.

Table 15 provides summary statistics for the final sample of 17 test durations. Similar to the initial evaluation durations, the SCV is less than 1. Figure 16 shows that a gamma distribution provides a reasonable fit to the empirical distribution, taking into consideration the small sample size. The maximum-likelihood parameter estimates for the gamma distribution are $\hat{k} = 2.363$ and $\hat{\theta} = 28.95$, resulting in a mean of $\hat{k}\hat{\theta} = 68.4$ minutes and SCV $= 1/\hat{k} = 0.423$ for the fitted distribution. We use this gamma distribution in the DES base case.

F. DES Model Details

We developed the DES model using the Arena software (Version 15.10.00001). All DES results are based on 250 replications of 30 days, excluding a 2-day warm-up period.

The model (global) variables are:

- $\hat{\theta}$: threshold value
- ψ: proportion of TT patients

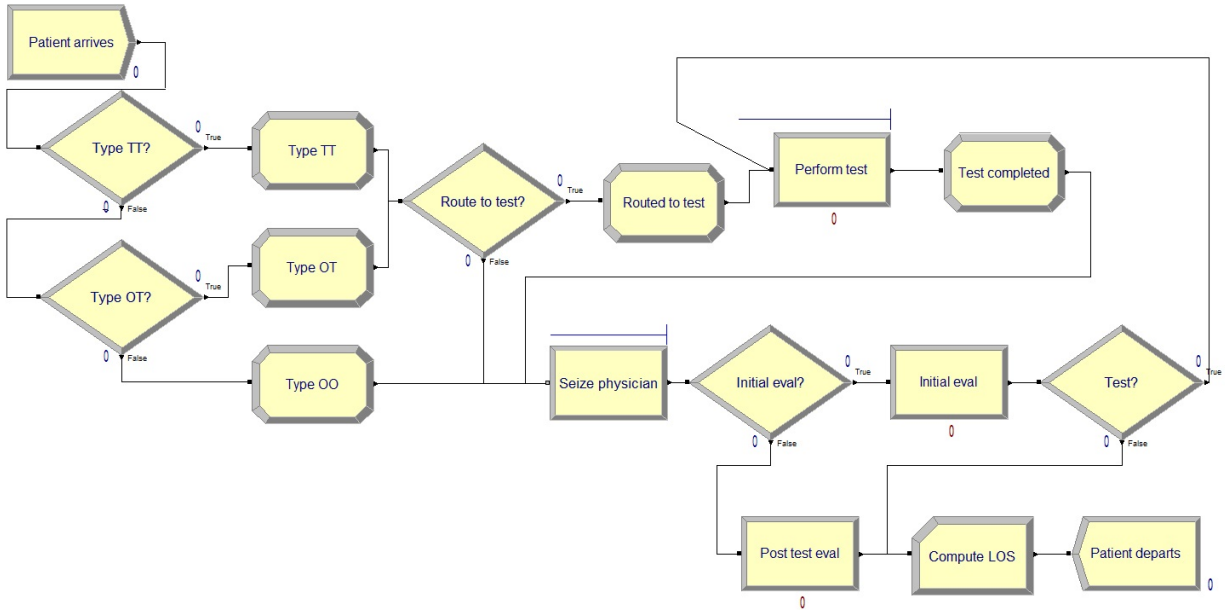

Figure 17 DES model Arena flowchart.

- ν: proportion of $O*$ patients triaged as OO
- vArrivalRate: an array of 24 hourly arrival rates
- vNumPhysicians: an array of 24 values for physician capacity
- κ_1 and θ_1: the shape and scale parameters for Gamma distribution for the initial evaluation time
- κ_2 and θ_2: the shape and scale parameters for Gamma distribution for test time
- κ_3 and θ_3: the shape and scale parameters for Gamma distribution for the post test evaluation time

The model has two resources: Physicians (with capacity that varies according to vNumPhysicians) and Test (with a capacity of 1).

The model entities are patients. The attributes of a Patient are:

- patientType: Initialized to TT, OT, or OO, with probabilities $\psi, (1-\nu)(1-\psi)$, or $\nu(1-\psi)$
- bTestCompleted: Initialized to FALSE
- bRoutedToTest: Initialized to FALSE
- $\tau_i, i = 1, 2, 3$: Initialized to a Gamma random variate with parameters κ_i and θ_i
- LOS: Computed immediately before the entity is disposed of

Figure 17 shows the Arena flowchart for the DES model. Patient entities are generated through the Patient arrives create module, according to an arrivals schedule that uses the hourly arrival rates in vArrivalRate. In Arena, an arrivals schedule generates arrivals from a

non-homogeneous Poisson process with piece-wise constant arrival rates. Next, the entity's attributes are initialized, including the patientType attribute.

Patients of type OO are routed to the Seize physician queue. Patients of types OT and TT are routed from the Route to test? decision block to the Perform test process module if the difference between the number of patients waiting for or receiving service from Physicians and the number of patients waiting for or receiving service from Test exceeds $\hat{\theta}$, and to the Seize physician queue otherwise. If a patient is routed from Route to test? to Perform test, then the patient's bRoutedToTest attribute is set to TRUE in the Routed to test assignment module.

A patient captures a Physicians resource according to the FCFS policy. If more than one physician is available, the patient entering service is randomly assigned to one of the free physicians.

After seizing a physician, the patient entity moves to the Initial eval? decision block. If the patient's bTestCompleted attribute equals TRUE, then the patient is routed to the Post test eval delay/release module, delaying the captured physician for the duration of the patient's τ_3 attribute. Otherwise, the patient is routed to the Initial eval delay/release module, delaying the captured physician for the duration of the patient's τ_1 attribute.

Upon leaving the Post test eval module, the patient releases the captured physician. The patient is then routed to the Compute LOS module, where the LOS attribute is calculated as the time interval between the current simulated time and the entity's creation time. After that, the entity is disposed of in the Patient departs module.

Upon leaving the Initial eval module, the patient releases the captured physician. The patient is then routed to the Test? decision block. Patients with a patientType of TT are routed to the Perform test queue. Patients of the OT and OO type are routed to the Compute LOS module, and then are disposed of.

Patients in the Perform test queue capture the Test resource in FCFS order. After capturing the resource, the patient holds the resource for the duration of the patient's τ_2 attribute. From the Perform test module, the patient is routed to the Test completed assignment module, where the bTestCompleted attribute is set to TRUE. The patient is then routed to the Seize physician queue.

Six separate random number streams are used for the following purposes: Generate the τ_1 attribute, generate the τ_2 attribute, generate the τ_3 attribute, generate the patientType attribute (two streams are used for this), and generate patient arrivals. A seventh random number stream is used for routing, in the randomized routing experiment.