

Complexity-Augmented Triage: A Tool for Improving Patient
Safety and Operational Efficiency

Wallace J. Hopp

Stephen M. Ross School of Business
at the University of Michigan

Jeffrey S. Desmond

Emergency Department
University of Michigan Hospital

Soroush Saghafian

Dept. of Industrial & Operational Engineering
University of Michigan

Mark P. Van Oyen

Dept. of Industrial & Operational Engineering
University of Michigan

Steven L. Kronick

Emergency Department
University of Michigan Hospital

Ross School of Business Working Paper

Working Paper No. 1161

Sept 2013

This work cannot be used without the author's permission.

This paper can be downloaded without charge from the
Social Sciences Research Network Electronic Paper Collection:

<http://ssrn.com/abstract=1911376>

Complexity-Augmented Triage: A Tool for Improving Patient Safety and Operational Efficiency

(Authors' names blinded for peer review)

Hospital Emergency Departments (ED's) typically use triage systems that classify and prioritize patients almost exclusively in terms of their need for timely care. We demonstrate that this is less effective than a triage system that adds an up-front estimate of patient complexity to the conventional urgency-based classification. Using analytic and simulation models calibrated with hospital data, we show that this complexity-augmented triage can substantially improve both patient safety (by reducing the risk of adverse events) and operational efficiency (by shortening the average length of stay). Moreover, we find that ED's with high resource (physician and/or examination room) utilization, high heterogeneity in the treatment time between simple and complex patients, and a relatively equal number of simple and complex patients benefit most from the complexity-augmented triage. Finally, we find that: (1) Whereas misclassification of a complex patient as simple is slightly more harmful than vice versa, complexity-augmented triage is relatively robust to misclassification error rates as high as 25%. (2) Streaming patients based on complexity information and prioritizing them based on urgency is better than doing the reverse. (3) Separating simple and complex patients via streaming facilitates the application of lean methods that can further amplify the benefit of complexity-augmented triage.

Key words: Healthcare Operations; Emergency Department; Triage; Priority Queues; Patient prioritization; Markov Decision Processes.

1. Introduction

Triage (a word derived from the French verb “trier,” meaning “to sort”) refers to the process of sorting and prioritizing patients for care. FitzGerald et al. (2010) noted that there are two main purposes for triage: “[1] to ensure that the patient receives the level and quality of care appropriate to clinical need (clinical justice) and [2] that departmental resources are most usefully applied (efficiency) to this end.” (see Moskop and Ierson (2007) for further discussion of the underlying principles and goals of triage).

While current triage systems used around the world address the clinical justice purpose of triage, the efficiency purpose has been largely overlooked. For instance, most ED's in Australia use the Australasian Triage Scale (ATS), the Manchester Triage Scale (MTS) is prevalent in the U.K., and ED's in Canada generally use the Canadian Triage Acuity Scale (CTAS). While they differ in their details, all of these triage systems classify patients strictly in terms of urgency and so address only the first (clinical justice) purpose of triage.

In the U.S., many ED's continue to use a traditional urgency-based 3-level triage scale, which categorizes patients into emergent, urgent, and non-urgent classes. But a growing majority of U.S. hospitals have adopted 5-level triage systems (see Fernandes et al. (2005)), which combine urgency with an estimate of resources (e.g., tests) required. In these systems (a typical version of which is illustrated in Figure 1 (left)), urgent patients who cannot wait are classified as level 1 or 2, while

non-urgent patients who can wait are classified as level 3, 4, or 5. Level 4 and 5 patients are usually directed to a fast track (FT) area, while level 1 patients are almost always moved immediately to a resuscitation unit (RU). Level 2 and 3 patients, who represent the majority of patients at large academic hospitals (about 80% at the University of Michigan Hospital ED (UMHED)), are served in the main area of the ED with priority given to level 2 patients. Since 5-level systems do not differentiate between level 2 and 3 patients in terms of complexity, patients in the main ED (about 80% of patients) are still sorted and prioritized purely on the basis of urgency. Hence, although 5-level triage systems represent an improvement over traditional 3-level triage scales, they remain urgency based systems for the majority of patients. In this paper we propose an augmented triage system, which we term *complexity-augmented* triage, that can significantly improve performance of the main ED with respect to both clinical justice and efficiency.

This poses two challenges: (a) deciding what information to collect at triage, and (b) determining how to use the information to improve performance. There are two main choices for the latter: prioritization and streaming. But they can be combined by using some information to separate patients into streams and some other information to prioritize them within the streams. This poses an additional question: what information to use to stream patients and what information to use to prioritize them?

Prioritization and streaming are not new. All ED's prioritize patients according to urgency. Many large ED's stream low acuity patients into fast tracks. But in recent years new types of streaming have received attention from both practitioners and researchers (see King et al. (2006), Ben-Tovim et al. (2008), and Saghafian et al. (2012)). Particularly relevant to this paper is our previous work in Saghafian et al. (2012), which showed that ED's can improve performance by having triage nurses predict the final disposition (admit or discharge) of patients and using this information in a "virtual streaming" patient flow design. That study showed that assigning patients to separate admit and discharge streams can reduce average time to first treatment for admit patients and average length of stay for discharge patients. But it also indicated that the performance of the streaming policy improves as the difference between the average treatment times of admit and discharge patients becomes larger. Since complexity is a better proxy for treatment time than is disposition, this suggests that classifying patients according to complexity may be even more useful than classifying them according to disposition.

Referring to procedures, investigations, or consultations as "interactions," we propose the new complexity-augmented triage process depicted in Figure 1 (right). Unlike a conventional 5-level system which makes no complexity distinction among the level 2 and 3 patients that make up the majority of ED patients, our proposed system systematically classifies them in terms of complexity. The additional step required in triage (i.e., predicting whether the patient will need two or

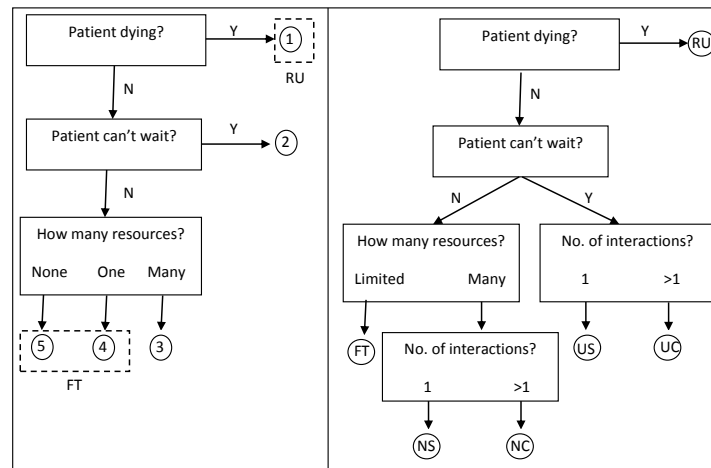


Figure 1 Left: Typical 5-level triage system (see, e.g., Gilboy et al. (2005)); Right: Proposed complexity-augmented triage system (RU: Resuscitation Unit, FT: Fast Track, NS: Non-urgent Simple, NC: Non-urgent Complex, US: Urgent Simple, UC: Urgent Complex).

more interactions) can be performed in seconds, and hence, does not add any significant amount of time to triage. However, it is unclear how much this additional information can improve the ED performance in terms of risk of adverse events (clinical justice) and average length of stay (efficiency), since it is subject to misclassification errors. To clarify this and related issues, we make use of a combination of analytic and simulation models calibrated with hospital data to address the following:

1. **Prioritization:** *How should ED's use complexity-augmented triage information to prioritize patients?*
2. **Magnitude:** *How much benefit does complexity-augmented triage (which adds complexity information to conventional urgency evaluations) offer relative to urgency-based triage?*
3. **Sensitivity:** *How sensitive are the benefits of complexity-augmented triage to misclassification errors and other characteristics that may vary across ED's?*
4. **(Patient Flow) Design:** *Is complexity information more effective if used to prioritize patients or to separate patients into streams?*

The main contribution of this paper is to provide insights of value to ED managers by addressing the above questions. However, the above questions also require addressing some technical challenges: (1) In the ED, upfront triage misclassifications are inevitable. We incorporate misclassifications through a linear transformation of control indices so that they represent “error-impacted” rates, which use only information from historical data. This leads to a modified version of the well-known $c\mu$ rule for the case with customer misclassification (in which control indices are replaced with their linearly transferred “error-impacted” counterparts). Furthermore, while these results are obtained by modeling the occurrence of adverse events as Poisson processes, in Online Appendix C, we use sample path arguments and appropriate notions of stochastic ordering to demonstrate the robustness of the priority rules under more general adverse event occurrence processes. (2) To provide guidance for ED physicians on how to prioritize patients within the examination rooms

(when they have a choice of what patient to see next), we develop a Markov Decision Process (MDP) model. A challenging feature of this model, which is common in many other health delivery settings, is that patients are sent for tests (e.g., MRI, CT Scan, X-Ray, etc.), and are unavailable to the physician during testing. In such a setting, the physician must consider both the current and the future availability of the patients when making decisions. This type of problem usually results in complex state-dependent optimal control policies. However, we show how a simple-to-implement rule that relies only on historical data defines the optimal policy for ED physicians. (3) Because of unbounded transition rates, the continuous MDP model of patient prioritization within examination rooms cannot use the conventional method of uniformization of Lippman (1975). There are very few results for continuous-time MDP's with unbounded transition rates (see, e.g., Guo and Liu (2001)). We contribute to this literature by showing how one can use a sequence of MDP's, each with bounded transition rates, to derive an optimal policy for the original MDP.

The remainder of the paper is organized as follows. Section 2 summarizes previous research relevant to our research questions. Section 3 describes our performance metrics and analytical modeling approach. For modeling purposes, we divide the ED experience of the patient into Phase 1 (from arrival until assignment to an examination room) and Phase 2 (from assignment to an examination room until discharge/admission to the hospital). Section 4 addresses Phase 2 by developing and analyzing a Markov Decision Process model. The result of the Phase 2 model is then used in Section 5, which focuses on Phase 1 and develops analytical queueing models to compare performance under urgency-based and complexity-augmented triage. Section 6 uses a realistic simulation model of the full ED calibrated with hospital data to validate the insights obtained through our analytical models and to refine our estimates of the magnitude of performance improvement possible with complexity-augmented triage. We conclude in Section 7.

2. Literature Review

The effect of assigning priorities in queueing systems has been well studied in the operations research literature. Analyzing a two-priority single-channel system, Cobham (1954, 1955) assumed perfect classification and van der Zee and Theil (1961) solved the case of imperfect classification. Under perfect classification, an average holding cost objective, Poisson arrivals, and a non-preemptive non-idling single server model, Cox and Smith (1961) showed that the $c\mu$ rule is optimal among priority rules. Kakalik and Little (1971) extended this result to show that the $c\mu$ rule remains optimal even among the larger class of state-dependent policies with or without the option of idling the server. The $c\mu$ rule has since been shown to be optimal in many other queueing frameworks; see, e.g., Buyukkoc et al. (1985), Van Mieghem (1995), Argon and Ziya (2009), Saghafian et al. (2011), and references therein.

Several studies that analyze patient flow in ED's from an operations perspective are related

to our work. Siddharathan et al. (1996) considered the impact of non-emergency patients on ED delays using urgency-based triage, and proposed a simple priority queueing model to reduce average waiting times. Wang (2004) considered a queue of heterogeneous high risk patients with perfect classification, and concluded that patients should be prioritized into as many urgency classes as possible in order to maximize survival. Argon and Ziya (2009) used average waiting time as the performance metric in a service system with two classes of customers, in which customer classification is imperfect, and showed that prioritizing customers according to the probability of being from the class that should have a higher priority when classification is perfect outperforms any finite-class priority policy. Dobson et al. (2013) developed a heavy-traffic model with an investigator and server interruptions to study physician choice in prioritizing patients. Huang et al. (2012) modeled the ED physician capacity as a queueing system with multi-class customers, where some customers need to be seen within a time period and others feed back through the service and have congestion related costs.

In the medical literature, Gilboy et al. (2005), FitzGerald et al. (2010), and Ierson and Moskop (2007) provide excellent reviews of the history of the triage process and its development over time. Although triage has been based mainly on urgency, the idea of considering the complexity of patients goes back to World War I mass-casualty triage recommendations: “*A single case, even if it urgently requires attention, –if this will absorb a long time,– may have to wait, for in that same time a dozen others, almost equally exigent, but requiring less time, might be cared for. The greatest good of the greatest number must be the rule.*” (Keen (1917)). Anticipating the potential of complexity-augmented triage, Vance and Spirvulis (2005) empirically tested the ability of nurses to estimate patient complexity at the time of triage and found that they are able to do this reliably. Vance and Spirvulis (2005) suggested that this type of information could be used to improve patient flow in ED’s, although they did not specify how. Finally, it is noteworthy that similar complexity information is also used in mass-casualty triage settings (see, e.g., Mills et al. (2013) and the references therein).

3. Modeling the ED

To address the four questions (prioritization, magnitude, sensitivity, and design) posed in Section 1, we developed a model of patient flow through the main ED (see Figure 2). We focus our attention on the main ED, which means we do not consider the minority of the patients routed to the resuscitation unit or fast track. A patient’s path through the main ED begins with *arrival*, which occurs in a non-stationary stochastic manner. Upon arrival, the patient goes to *triage*, where s/he is classified according to a predefined process (based on urgency and/or complexity), which inevitably involves some misclassification errors. If an examination room is not immediately available, s/he goes to the *waiting* area until s/he is called by the charge nurse and brought to an examination

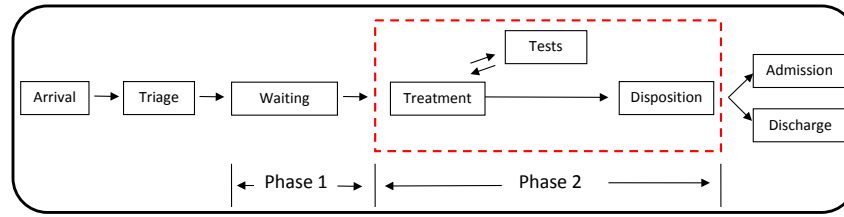


Figure 2 General flow of patients in the main ED.

room. There s/he goes through a stochastic number of *treatment* stages with a physician, which include diagnosis, consultation and other interactions that are also stochastic in duration. These treatment stages are punctuated by *test* stages during which the patient is unavailable to the physician, which involve testing (MRI, CT Scan, X-Ray etc.), preparation/processing activities that do not involve the physician, or waiting for test results. The final processing stage after the last physician interaction is *disposition*, in which the patient is either *discharged* to go home or *admitted* to the hospital.

We refer to the time a patient spends after s/he is triaged and before s/he is brought an examination room as “Phase 1,” and label the remaining time until disposition as “Phase 2.” Because they are under observation and care, patients have a lower risk of adverse events during Phase 2 than during Phase 1. Patients are taken from Phase 1 to Phase 2 by the charge nurse based on a Phase 1 sequencing rule that uses the patient classification performed at triage. Similarly, in Phase 2, physicians use some kind of sequencing rule to choose which patient to see next.

During the patient’s stay in the ED, s/he may experience adverse events, which we define to be degradations in health status that are associated with worse outcomes (e.g., Brennan et al. (1991) and Diercks et al. (2007)). There are various examples of such events including rectal bleeding (Appendix I of Brennan et al. (1991)), chest compression, hypertension, tachyarrhythmia, and bradyarrhythmia (Scheuermeyer et al. (2010)) among others. A patient may experience more than one adverse event unless, of course, the event is death. But because death is so rare relative to the rate of adverse events (e.g., Liu et al. (2005) report that 28% of patients boarded in the ED experienced some type of adverse event (including errors), while Baker and Clancy (2006) reported an ED death rate of 0.26%), we do not include it as a terminating adverse event. Furthermore, because most adverse events are not visible to providers at the time they occur (e.g., Brennan et al. (1991)), we do not allow patient priorities to be reassigned as a result of them.

It is widely known that a higher waiting time is associated with a higher risk of adverse events (e.g., Diercks et al. (2007) report that patients with a longer ED time have a higher risk of recurrent myocardial infarction). We model this effect by representing the occurrence of adverse events with type dependent Poisson processes. However, we relax the Poisson assumption in Online Appendix C, and allow the processes approximating the occurrence of adverse events to be any general stationary point process.

In our framework, a patient’s rate of adverse events is influenced by his/her true type

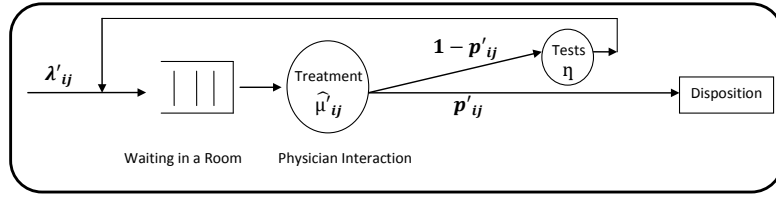


Figure 3 Patient flow after a patient is moved to an examination room/bed (Phase 2 sequencing).

$ij \in \mathcal{U} \times \mathcal{C}$, where $i \in \mathcal{U}$ is his/her urgency level, $j \in \mathcal{C}$ is his/her complexity type, $\mathcal{U} = \{U(\text{Urgent}), N(\text{Non-urgent})\}$ and $\mathcal{C} = \{C(\text{Complex}), S(\text{Simple})\}$. Under urgency-based triage an estimate is made of i , while under complexity-augmented triage estimates are made of both i and j .

Since sequencing decisions in Phase 1 may depend on patients' ED service times (the time they spend in Phase 2), they may be affected by Phase 2 prioritization. Thus, we start by analyzing Phase 2 and then use our Phase 2 results to justify a model with which to derive an optimal Phase 1 sequencing rule. Finally, we test the insights gained from our analytic models under realistic conditions with a simulation model of the full ED calibrated with a year of data from the University of Michigan Hospital ED (UMHED) as well as time study data from the literature.

4. Phase 2: Sequencing Patients Within the ED

To model Phase 2, we consider the multi-stage service process illustrated in Figure 3. We start by considering the system under the assumption of exogenous arrivals to Phase 2. This situation occurs in practice during periods when the ED has sufficient bed and physician capacity to allow patients to move almost immediately into the examination rooms without waiting in the ED. For tractability, we assume patients *classified* as $ij \in \mathcal{U} \times \mathcal{C}$ arrive according to a Poisson process with rate λ'_{ij} . Note that we use superscript “ ’ ” throughout the paper to indicate error-impacted rates. Such rates can be directly estimated from arrival data after patients are classified, but we will also provide in Sections 5.1 and 5.2 a way to calculate them using the raw arrival rates. We assume patients of type ij are subject to adverse events which occur according to a Poisson process. We denote the error-impacted intensity of adverse events in Phase 2 by the vector $\hat{\theta}' = (\hat{\theta}'_{ij})_{ij \in \mathcal{U} \times \mathcal{C}}$ (which we expect to be less than that in Phase 1, denoted by θ' , because of monitoring and treatment patients receive in the examination rooms). As they enter examination rooms, patients are assigned to physicians who treat them, often with multiple visits, until their discharge or admission to the hospital. Since an individual physician may be assigned to several patients s/he often has a choice about who to see next among his/her available patients. To construct a simplified analytic model, we aggregate preparation time, test time, and waiting time for the test results. Patients who have completed a test or tests ordered by the physician, and have all of the associated results ready, are termed “available” for a physician visit, while patients being tested, prepared, or waiting for results are labeled “unavailable.”

Letting $R_{\pi}^{\Omega}(t)$ represent the counting process that tallies the total number of adverse events (for

all patients) until time t under patient classification (triage) policy Ω and sequencing rule π , we define $R_\pi^\Omega = \lim_{t \rightarrow \infty} R_\pi^\Omega(t)/t$ (when the limit exists) as our metric and refer to it as the *Rate of Adverse Events (ROAE)*. However, if $\hat{\theta}_{ij} = 1$ for all $i \in \mathcal{U}$ and $j \in \mathcal{C}$, then it can be shown that $R_\pi^\Omega / \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{C}} \hat{\theta}_{ij} \lambda_{ij} = R_\pi^\Omega / \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{C}} \lambda_{ij}$ reduces to average *Length of Stay (LOS)*. (Notice that the sample path costs of LOS and that of adverse events under unit risk intensities divided by the total arrival rate will be different, but they are equal in expectation.) Hence, this observation allows us to use our metric to characterize performance with respect to both safety and efficiency.

For tractability, we assume each interaction with patients *classified* as ij takes an exponentially distributed amount of time with rate $\hat{\mu}'_{ij}$. We also assume that the physician can preempt an interaction to see a patient of a different class. When a physician returns to a preempted interaction, we assume s/he must repeat the process (e.g., review vital signs, lab results, etc.), and so we assume a preempt-repeat protocol. In practice, emergency physicians can, and sometimes do, preempt patients to deal with emergencies. But for fairness and efficiency reasons, they do this rarely. Hence, we test our conclusions under the assumption of non-preemption in Phase 2 in Section 6 using simulation. After each completed interaction, a patient classified as ij may be disposed (discharged home or admitted to the hospital) with probability $p'_{ij} > 0$, or with probability $1 - p'_{ij}$ requires another round of test and treatment. We note that in practice the probability of being disposed may not be constant because it depends on various factors (e.g., progression of pain, the number of past interactions with the physician, revealed test results, etc.). If data on such factors were collected, they could be incorporated into the patient prioritization decision. Since such data do not currently exist, we approximate the number of interactions with the physician by fitting a geometric distribution with constant probability of departure p'_{ij} for patients classified as $ij \in \mathcal{U} \times \mathcal{C}$, which can easily be estimated using the average number of physician-patient interactions for that class. Empirical data from the literature (Graff et al. (1993)), shown in Figure 5, suggest that this is a reasonable approximation of reality.

We model the aggregated test times (including any preparation and waits for results) as i.i.d. exponential random variables. This is a justifiable approximation because: (a) most testing facilities are shared with units outside the ED and so have workloads not visible to ED personnel, and (b) even if test facility workloads were known, ED personnel could not incorporate them into patient sequencing decisions because they do not know which tests, if any, will be required before examining a patient and/or his/her prior test results (which begins an interaction). So, for purposes of patient sequencing at least, test time delays look like i.i.d. random variables to the physician. Modeling such delays as exponential is reasonable because (1) the waiting time distribution in many queueing systems is exponential or nearly so, and (2) the physician does not keep track of the “age” of such delays. (For more detailed discussions regarding test/diagnostic facilities, we refer

interested readers to Green et al. (2006a), Patrick et al. (2008), and Batt and Terwiesch (2012).) To construct a tractable analytic model, the aggregate test delay times are further assumed to be a generic “test” with mean time η^{-1} that is the same across different patient classes. However, we relax this assumption and allow patient class specific test delays in our simulation model of Section 6. Finally, we note that in most ED’s physicians do not update patients’ triage classes for various reasons including those related to liability. Hence, consistent with practice, we assume patient classifications are made at triage and are not updated during the Phase 2 service process. We refer to the representation of Phase 2 of the ED service with above assumptions as the *simplified Phase 2 model with dynamic arrivals*.

Because each physician is dedicated to his/her own slate of patients, we focus on a single physician’s decision of who to see next. To this end, we let $\underline{x} = (x_{ij})_{ij \in \mathbf{u} \times \mathbf{c}}$ (respectively $\underline{y} = (y_{ij})_{ij \in \mathbf{u} \times \mathbf{c}}$) represent the error-impacted number of patients of each class available (not available) for the physician visit. With these, we can define the state of the system at any point in time, t , by the vector $(\underline{x}(t), \underline{y}(t)) \in \mathbb{Z}_+^4 \times \mathbb{Z}_+^4$, and model the process $\{(\underline{x}(t), \underline{y}(t)) : t \geq 0\}$ as a Continuous Time Markov Chain (CTMC). We assume the parameters of the system are such that this CTMC is stabilizable; that is, there exists at least one policy under which the risk of adverse events is finite (otherwise, the problem does not represent a real ED). However, notice that we model test delays as a $\cdot/M/\infty$ queueing system. Since transition rates are not bounded in our framework, we cannot use the uniformization method of Lippman (1975) to formulate a discrete time equivalent of the CTMC in which the times between consecutive events are i.i.d. (for all states). So instead, we construct a sequence of Controlled CTMC’s (CCTMC’s) with an increasing but bounded sequence of (maximum) transition rates converging to the original CCTMC. We do this by replacing the $\cdot/M/\infty$ test stage with four parallel $\cdot/M/k$ systems (one devoted to each patient class), index the underlying CCTMC with k , and let $k \rightarrow \infty$. The advantage of having four parallel $\cdot/M/k$ queues (instead of one $\cdot/M/k$) is that the order of jobs in each queue becomes irrelevant, and hence, does not need to be captured in the system’s state. Another novel aspect of our approach is that we truncate the transition rates instead of truncating the state space, thereby avoiding the artificial boundary effects that usually distort the optimal policy. Since the transition rates in the CTMC indexed by k (for all k) are bounded by $\psi_k = \max_{ij \in \mathbf{u} \times \mathbf{c}} \hat{\mu}'_{ij} + 4k\eta + \sum_{ij \in \mathbf{u} \times \mathbf{c}} \lambda'_{ij} < \infty$, we can use the standard uniformization technique to derive the optimal policy for each CCTMC. We then use a convergence argument (taking the limit as $k \rightarrow \infty$) to derive the optimal policy for the original problem. It should be noted that we can always start with a sufficiently large k such that the stability of the underlying system is not affected (since the original system is stable by assumption).

For the system indexed by k , the optimal rate of adverse events under a patient classification based on both sets \mathbf{u} and \mathbf{c} , $R^{k*} = \inf_{\pi \in \Pi} R_{\pi}^{\mathbf{u} \cup \mathbf{c}}$ (where Π denotes the set of all preemptive admis-

sible Markovian policies), and the optimal physician behavior can be derived from the following average cost optimality equation:

$$\begin{aligned} J^k(\underline{x}, \underline{y}) + R^{k*} = & \frac{1}{\psi_k} \left[\hat{\theta}'(\underline{x} + \underline{y})^T + \sum_{ij \in \mathbf{U} \times \mathbf{C}} [\lambda'_{ij} J^k(\underline{x} + \underline{e}_{ij}, \underline{y}) + (y_{ij} \wedge k) \eta J^k(\underline{x} + \underline{e}_{ij}, \underline{y} - \underline{e}_{ij})] \right. \\ & + \min_{a \in \mathcal{A}(\underline{x})} \left\{ \sum_{ij \in \mathbf{U} \times \mathbf{C}} \mathbb{1}_{\{a=ij\}} \hat{\mu}'_{ij} [p'_{ij} J^k(\underline{x} - \underline{e}_{ij}, \underline{y}) + (1 - p'_{ij}) J^k(\underline{x} - \underline{e}_{ij}, \underline{y} + \underline{e}_{ij})] \right. \\ & \left. \left. + (\psi_k - \sum_{ij \in \mathbf{U} \times \mathbf{C}} [\lambda'_{ij} + (y_{ij} \wedge k) \eta + \mathbb{1}_{\{a=ij\}} \hat{\mu}'_{ij}]) J^k(\underline{x}, \underline{y}) \right\} \right], \quad (1) \end{aligned}$$

where $J^k(\underline{x}, \underline{y})$ is a relative cost function (defined as the difference between the total expected cost of starting from state $(\underline{x}, \underline{y})$ and that from an arbitrary state such as $(\underline{0}, \underline{0})$), $a \wedge b = \min\{a, b\}$, \underline{e}_{ij} is a vector with the same size as \underline{x} with a 1 in position ij and zeroes elsewhere, a is an action determining which patient class to serve, and $\mathcal{A}(\underline{x}) = \{ij \in \mathbf{U} \times \mathbf{C} : x_{ij} > 0\} \cup \{0\}$ is the set of feasible actions (class 0 represents the idling action) when the error-impacted number of patients of each class in the examination rooms is \underline{x} .

Although our model of Phase 2 has a complex multi-stage structure with feedback (i.e., random patient returns after each visit), which generally makes the optimal policy complex (see, e.g., Tcha and Pliska (1977)), the optimal behavior of the physician can be described by an appealingly simple operational rule (proofs of this and all other results are given in Online Appendix A).

THEOREM 1 (Phase 2 Prioritization). *In the simplified Phase 2 model with dynamic arrivals, regardless of the number and class of available and unavailable patients, the physician should prioritize available patients in decreasing order of $p'_{ij} \hat{\theta}'_{ij} \hat{\mu}'_{ij}$. Furthermore, the physician should not idle when there is a patient available in an exam room.*

The prioritization index in Theorem 1 is computed as the probability that the visit will be the final interaction with the patient (p'_{ij}) times the estimated risk of adverse events ($\hat{\theta}'_{ij}$) divided by the average duration of each visit ($1/\hat{\mu}'_{ij}$). Such a policy is easy to implement, since (a) the physician does not need to consider the number and class of patients available in the examination rooms or under tests, and (b) the physician (or decision support system) can easily estimate the required quantities. (For example, the authors have developed a smart phone application that can be used by ED physicians to facilitate collection of required data and computation of patient priorities.) In most settings, θ'_{ij} is larger when i is U (urgent) than when i is N (non-urgent), and p'_{ij} and $\hat{\mu}'_{ij}$ are larger when j is S (simple) than when j is C (complex). Since the relative difference in θ'_{ij} is much larger than the relative difference in p'_{ij} and $\hat{\mu}'_{ij}$, it follows that: US (Urgent Simple), UC (Urgent Complex), NS (Non-urgent Simple), NC (Non-urgent Complex) defines the optimal Phase 2 priority policy for most hospitals.

As a further check on the robustness of this prioritization result, we consider an alternate model in which arrivals to Phase 2 are purely endogenous. Specifically, we assume that all patients for a

given time interval (e.g., a day) arrive at once and the objective is to clear them out as quickly as possible to minimize LOS and ROAE. We further assume that ED physicians can treat any patient in the system (i.e., there are no constraints on the number of beds or the number of patients per physician), so that patients are drawn into Phase 2 entirely through the sequencing decisions of the physicians. We label this the simplified Phase 2 model with static arrivals. In contrast with the simplified Phase 2 model with dynamic arrivals, which is representative of the ED under light load conditions, this model is representative of the ED under heavy load conditions that create a backlog of patients. In Online Appendix D we show that the Phase 2 sequencing policy of Theorem 1 remains optimal under these very different modeling conditions. The suggestion is that the cost/time balance struck by the $c\mu$ -type rule of Theorem 1 is robustly effective in Phase 2.

However, in practice, the ED oscillates between underload and overload conditions. Also, constraints on the number of beds and the number of patients per physician sometimes prevent idle physicians from taking a new patient. Both of these realities make the interface between Phase 1 and Phase 2 more complex than in either of simplified models considered here. To find an effective way to manage this interface and to see whether the Phase 2 sequencing rule remains effective in realistic settings, we proceed in two steps. First, we examine a simplified model of Phase 1 to gain insights into optimal sequencing of patients into the ED. Second, we use a realistic simulation of the combined ED to determine whether the policies suggested by the simplified models are effective in the actual system.

5. Phase 1: Sequencing Patients Into the ED

To create a simplified model that captures the essential dynamics of sequencing patients into the ED, we represent the dashed area in Figure 2 (i.e., Phase 2) as a single-stage aggregated service node with a single “super server” that represents the aggregate ED capacity. Because our Phase 2 analysis indicated that simple patients with a given urgency level should be prioritized over complex ones within the ED and, by definition, complex patients have on average more interactions with physicians it follows that simple patients should have a higher aggregate/effective service rate than complex ones in this simplified model. Specifically, we suppose patients of type $ij \in \mathcal{U} \times \mathcal{C}$ have i.i.d. service times (i.e., the total time spent in Phase 2) that follow a general distribution, $F_{ij}(s)$ with first moment $1/\mu_{ij}$, where $\mu_{iC} \leq \mu_{iS}$ for all $i \in \mathcal{U}$, and a finite second moment. For tractability, we model the arrival of patients of type $ij \in \mathcal{U} \times \mathcal{C}$ to the ED as a Poisson process with rate λ_{ij} . Moreover, as we did in our simplified Phase 2 models, we assume patients of type ij are subject to adverse events, which occur according to a Poisson process with intensity θ_{ij} , where $\theta_{Uj} \geq \theta_{Nj}$ for all $j \in \mathcal{C}$. These assumptions lead to a tractable model of Phase 1; however, many of them will be relaxed later (see, e.g., Online Appendix C and Section 6).

5.1. Urgency-Based Triage

We first consider current practice in most ED's in which patients in the main ED (levels 2 and 3) are classified solely based on urgency, and use our simplified Phase 1 model to examine decisions for sequencing patients into the ED. We start with the case of perfect classification and then consider the case of stochastic misclassification.

When patients can be perfectly classified as either urgent (U) or non-urgent (N), the arrival rates for U's and N's are $\lambda_U = \sum_{j \in \mathcal{C}} \lambda_{Uj}$ and $\lambda_N = \sum_{j \in \mathcal{C}} \lambda_{Nj}$, respectively. Similarly, the average service times for U's and N's are $1/\mu_U = \sum_{j \in \mathcal{C}} (\lambda_{Uj}/\lambda_U)(1/\mu_{Uj})$ and $1/\mu_N = \sum_{j \in \mathcal{C}} (\lambda_{Nj}/\lambda_N)(1/\mu_{Nj})$, respectively. Furthermore, from known results for non-preemptive priority queues (see, e.g., Cobham (1954)) the average waiting (queue) time of the k th priority class is

$$W_k = \frac{\lambda \mathbb{E}(s^2)}{2(1 - \sum_{l < k} \rho_l)(1 - \sum_{l \leq k} \rho_l)}, \quad (2)$$

where s represents the service time of a randomly chosen patient, and $\rho_l = \lambda_l/\mu_l$ for class l . Hence, if U's are prioritized over N's, then the average waiting time is $W_U = \lambda \mathbb{E}(s^2)/2(1 - \rho_U)$ for U's and $W_N = \lambda \mathbb{E}(s^2)/2(1 - \rho_U)(1 - \rho)$ for N's. Furthermore, the average intensity of adverse events for U's is $\theta_U = (\lambda_{US}/\lambda_U)\theta_{US} + (\lambda_{UC}/\lambda_U)\theta_{UC}$ and for N's is $\theta_N = (\lambda_{NS}/\lambda_N)\theta_{NS} + (\lambda_{NC}/\lambda_N)\theta_{NC}$. With these, the ROAE under an urgency-based triage policy (i.e., classification with respect to set \mathcal{U}) that gives priority to U's (denoted by $R_U^{\mathcal{U}}$) or N's (denoted by $R_N^{\mathcal{U}}$) follows:

$$R_U^{\mathcal{U}} = \theta_U \lambda_U (\lambda \mathbb{E}(s^2)/2(1 - \rho_U)) + \theta_N \lambda_N (\lambda \mathbb{E}(s^2)/2(1 - \rho_U)(1 - \rho)), \quad (3)$$

$$R_N^{\mathcal{U}} = \theta_N \lambda_N (\lambda \mathbb{E}(s^2)/2(1 - \rho_N)) + \theta_U \lambda_U (\lambda \mathbb{E}(s^2)/2(1 - \rho_N)(1 - \rho)). \quad (4)$$

Comparing these reveals that, without misclassification errors, the best priority rule is to prioritize U's (N's) if, and only if, $\theta_U \mu_U \geq (\leq) \theta_N \mu_N$. Given the criteria used to classify a patient as urgent, we expect θ_U and θ_N be such that $\theta_U \mu_U > \theta_N \mu_N$, meaning that U's will be given priority. However, this simple result may or may not hold if one considers the effect of stochastic triage misclassifications.

Therefore, we now formally incorporate stochastic misclassification errors into our model. Let γ_U and γ_N denote the misclassification probabilities for urgent and non-urgent patients, respectively. The arrival rates for patients classified (correctly or erroneously) as U and N are $\lambda'_U = \lambda_U(1 - \gamma_U) + \lambda_N \gamma_N$ and $\lambda'_N = \lambda_N(1 - \gamma_N) + \lambda_U \gamma_U$, respectively. Similarly, the mean service times for patients classified as U and N are $1/\mu'_U = [\lambda_U(1 - \gamma_U)(1/\mu_U) + \lambda_N \gamma_N(1/\mu_N)]/\lambda'_U$ and $1/\mu'_N = [\lambda_N(1 - \gamma_N)(1/\mu_N) + \lambda_U \gamma_U(1/\mu_U)]/\lambda'_N$, respectively. Finally, the intensity of adverse events for patients classified as U and N are $\theta'_U = [\lambda_U(1 - \gamma_U)\theta_U + \lambda_N \gamma_N \theta_N]/\lambda'_U$ and $\theta'_N = [\lambda_N(1 - \gamma_N)\theta_N + \lambda_U \gamma_U \theta_U]/\lambda'_N$, respectively.

Using (3) with these new error-impacted rates shows that when priority is given to U's, the ROAE under imperfect classification is:

$$R_U^{\mathcal{U}'} = \theta'_U \lambda'_U (\lambda \mathbb{E}(s^2)/2(1 - \rho'_U)) + \theta'_N \lambda'_N (\lambda \mathbb{E}(s^2)/2(1 - \rho'_U)(1 - \rho)), \quad (5)$$

where $\rho'_U = \lambda'_U/\mu'_U$. Similarly, using (4) shows that when priority is given to N's:

$$R_N^{\mathcal{U}'} = \theta'_N \lambda'_N (\lambda \mathbb{E}(s^2)/2(1 - \rho'_N)) + \theta'_U \lambda'_U (\lambda \mathbb{E}(s^2)/2(1 - \rho'_N)(1 - \rho)), \quad (6)$$

where $\rho'_N = \lambda'_N/\mu'_N$.

We summarize the implications of these results in the following proposition. Note that part (i) of this proposition coincides with the “expected $c\mu$ ” and Highest Signal First (HSF) policy discussed in Argon and Ziya (2009). However, due to differences between our setting and theirs (most importantly, Argon and Ziya (2009) assume a continuous signal from customers that indicates the probability of misclassification, while our approach uses average misclassification error rates that can be estimated from data), we provide an independent proof based on the above results.

PROPOSITION 1 (Phase 1 Prioritization - Urgency-Based Triage). *In the simplified Phase 1 model with imperfect urgency-based classification: (i) The best (static) priority rule is to prioritize U patients if $\theta'_U \mu'_U \geq \theta'_N \mu'_N$; otherwise, prioritize N patients. (ii) The best (static) priority rule is the same as that for the case without misclassification error if $\gamma_N + \gamma_U \leq 1$; otherwise, the best priority ordering is reversed.*

Part (i) of Proposition 1 shows how triage misclassification errors can be incorporated into the optimal priority rule. Part (ii) of Proposition 1 shows that if the misclassification rate is high enough, the optimal priority rule prioritizes N patients. However, empirical studies have observed misclassification levels γ_N and γ_U to be in the range 9-15% depending on the level of triage nurse experience (Hay et al. (2001)). Thus, if, as we expect, prioritizing urgent patients is optimal when there is no misclassification error, prioritizing them remains optimal even under practical levels of misclassification errors. This confirms that prioritizing level 2 patients over level 3 patients, as is typically done in the main ED, is reasonable. However, we note that there is wide variance of complexity among level 2 and level 3 patients (see, e.g., Vance and Spirvulis (2005) and Figure 4). Simply prioritizing level 2 patients over level 3 patients may be significantly suboptimal relative to a policy that considers complexity. We investigate this issue in the next section.

5.2. Complexity-Augmented Triage

We now consider the complexity-augmented triage policy shown in Figure 1 (right), and compare its performance to that of conventional urgency-based triage. By doing this we address the prioritization, magnitude, and sensitivity questions posed in the Introduction.

To evaluate the performance of complexity-augmented triage when classification is imperfect, we again let γ_U and γ_N denote the misclassification error rates with respect to set \mathcal{U} . Similarly, we let γ_C and γ_S denote the misclassification error rates with respect to set \mathcal{C} , γ_C denote the probability

that a C patient is classified as an S , and γ_S denote the probability that an S patient is classified as a C . We assume the misclassification probabilities with respect to \mathbf{U} and \mathbf{C} are independent because (a) the data show that the assessments themselves are uncorrelated, indicating that urgency and complexity are medically separable questions (e.g., Figure 4 indicates that an urgent patient is almost equal likely to be simple or complex (and vice versa)), and (b) multiple nurses perform triage, thereby limiting the extent of any systematic biases in misclassifications.

As noted earlier, misclassification error rates in terms of urgency have been observed to be in the range of 9-15% (Hay et al. (2001)). Vance and Spirvulis (2005) tested the ability of triage nurses to evaluate patient complexity and observed a misclassification rate of 17% (see also Kronick and Desmond (2009) for related empirical work in UMHED regarding the ability of triage nurses to classify patients).

Similar to what we did in Section 5.1, we need to calculate the error impacted rates λ'_{ij} , θ'_{ij} , and μ'_{ij} . Let $\underline{\lambda} = (\lambda_{US}, \lambda_{UC}, \lambda_{NS}, \lambda_{NC})$ and $\underline{\lambda}' = (\lambda'_{US}, \lambda'_{UC}, \lambda'_{NS}, \lambda'_{NC})$. Then $\underline{\lambda}'$ can be obtained through a linear transformation of $\underline{\lambda}$; $\underline{\lambda}'^T = A \underline{\lambda}^T$, where A is a (known) *misclassification error matrix*, and is defined as

$$A = \begin{pmatrix} (1 - \gamma_U)(1 - \gamma_S) & (1 - \gamma_U)\gamma_C & \gamma_N(1 - \gamma_S) & \gamma_N\gamma_C \\ (1 - \gamma_U)\gamma_S & (1 - \gamma_U)(1 - \gamma_C) & \gamma_N\gamma_S & \gamma_N(1 - \gamma_C) \\ \gamma_U(1 - \gamma_S) & \gamma_U\gamma_C & (1 - \gamma_N)(1 - \gamma_S) & (1 - \gamma_N)\gamma_C \\ \gamma_U\gamma_S & \gamma_U(1 - \gamma_C) & (1 - \gamma_N)\gamma_S & (1 - \gamma_N)(1 - \gamma_C) \end{pmatrix}. \quad (7)$$

Similarly, if $\underline{\theta}'$ and $\underline{\mu}'$ denote the vector of error impacted adverse event and service rates, we have $\underline{\theta}'^T = (A(\underline{\lambda} \times \underline{\theta})^T)/\underline{\lambda}'$ and $(\underline{1}/\underline{\mu}')^T = (A(\underline{\lambda}/\underline{\mu})^T)/\underline{\lambda}'$, where $\underline{1} = (1, 1, 1, 1)$ and operators “ \times ” and “/” are componentwise multiplication and division, respectively.

With these, the waiting times for each customer class under an imperfect $\mathbf{U} \cup \mathbf{C}$ classification can be computed using (2) with rates replaced with their transformed error impacted counterparts. This model permits us to show the following.

PROPOSITION 2 (Phase 1 Prioritization - Complexity-Augmented Triage). *In the simplified Phase 1 model with imperfect urgency and complexity classifications: (i) The best priority rule is to prioritize patients in decreasing order of $\theta'_{ij} \mu'_{ij}$ values. (ii) $R_*^{\mathbf{U}' \cup \mathbf{C}'} \leq R_*^{\mathbf{U}'}$. That is, even with misclassification errors, implementing the best priority rule for complexity-augmented triage is always (weakly) better than the optimal priority rule for urgency-based triage. (iii) The rule of part (i) is optimal even among the larger class of all non-anticipative (state or history dependent, idling or non-idling, etc.) policies.*

Proposition 2 (i) addresses the prioritization question by suggesting a simple priority rule analogous to the well-known “ $c\mu$ ” rule to incorporate complexity information into Phase 1 sequencing. However, the indices used are linearly transformed to incorporate misclassifications. Thus, this result can be viewed as an extension of the $c\mu$ rule under imperfect information. Furthermore,

Proposition 2 (i) shows precisely when the optimal priority rule will be different from the optimal rule without misclassification. For instance, if misclassification rates are high enough, it can be better to prioritize non-urgent patients over urgent ones. However, at the error levels observed in the previously cited studies, the implication of Proposition 2 (i) is to prioritize patients in the order: US, UC, NS, NC, which coincides with the priority rule we found to be optimal in Phase 2. Proposition 2 (ii) begins to address the magnitude question raised in the Introduction by suggesting that complexity-augmented triage outperforms urgency-based triage, provided that the optimal priority rule is implemented. While this result may seem intuitive due to the additional information collected at triage, we note that the additional information is subject to errors, so this conclusion is not obvious. Nevertheless, Proposition 2 (ii) shows that, implemented correctly, imperfect complexity-augmented information *always* improves the ED performance regardless of the misclassification levels. Moreover, it should be noted that information collection involves only simple estimation of whether two or more interactions are needed with a physician, which adds no appreciable time to the triage process. While priority rules are greedy and usually suboptimal, Proposition 2 (iii) confirms that they are optimal in this setting. The surprise is that it is never optimal to idle when only low priority patients are available, even though the model disallows preemption. (For instance, if preemption is not allowed and there is no US patient waiting to be seen, one might (incorrectly) expect it to be optimal for the physician to intentionally idle instead of starting to work on a patient with a long service time and a low urgency level (e.g., NC) in anticipation of arrival of a patient of US type.) Similar results for the $c\mu$ rule but without misclassifications are presented in Kakalik and Little (1971). Finally, part (iii) of Proposition 2 states that a dynamic (i.e., state-dependent) priority policy cannot beat the greedy and simple state-independent policy presented in part (i).

We can also address the sensitivity question by using our model to determine the environmental factors that favor complexity-augmented triage. We summarize the results in the following proposition:

PROPOSITION 3 (Attractiveness of Complexity-Augmented Triage). *Under the simplified Phase 1 model, the benefit of complexity-augmented triage compared to urgency-based triage (under their respected optimal priority policies), $R_*^{\mathbf{u}'} - R_*^{\mathbf{u}' \cup \mathbf{c}'}$, is (i) non-decreasing in ρ , (ii) non-decreasing in $1/\mu_C - 1/\mu_S$, (iii) maximized at $\alpha = 1/2$, when $\lambda_{US} = (1 - \alpha)\lambda_U$, $\lambda_{UC} = \alpha\lambda_U$, $\lambda_{NC} = \alpha\lambda_N$, and $\lambda_{NS} = (1 - \alpha)\lambda_N$, and (iv) non-increasing in γ_S and γ_C .*

This implies that complexity-augmented triage is more beneficial in ED's with (i) higher utilization, (ii) higher heterogeneity in the average service time of simple and complex patients, (iii) more equal fractions of simple and complex patients, and (iv) lower error rates in classifying simple and complex patients.

5.3. Patient Flow Design Using Complexity and Urgency Information

In this section, we examine the “design” question from the Introduction. In particular, we examine whether complexity information obtained at triage is more useful for separating patients into streams or for prioritizing them within streams. Additional insights will be provided in Section 6.4, where we use hospital data to compare the performance of different ED patient flow designs.

We consider two patient flow designs. In *complexity streaming*, S and C patients are sent to separate streams in which they are prioritized based on their urgency level (U before N). In *urgency streaming*, U and N patients are sent to separate streams, within which S patients are prioritized over C patients (consistent with the optimal priority rule established in Proposition 2 (i)). To make a fair comparison of these designs, we first remove the effect of unbalanced utilizations and assume that the ED can assign appropriate capacity (physicians, staff, beds, etc.) to streams so that their utilization becomes equal. We further assume that two conditions hold: (1) the (error impacted) effective mean service rates in each stream are equal, and (2) the variance of service times in each stream are equal. Since ROAE is a function of arrival rate, service rate (and hence utilization), as well as the second moment of service time (see Eq. 5), these assumptions provide a fair basis for comparing different streaming designs, which we term as *perfectly balanced* streaming designs, because they eliminate obvious differences in utilization-induced congestion that can be removed by appropriate capacity allocation between streams in each patient flow design. However, in Online Appendix B, we compare the performance of *partially balanced* streaming designs by relaxing these conditions, and we observe that our conclusions are robust. Finally, in Section 6.4, we use hospital data and simulation to further examine the performance of complexity-based streaming and the robustness of our conclusions.

PROPOSITION 4 (Patient Flow Design). *In perfectly balanced streaming systems, with each stream using its optimal policy suggested by Proposition 2 (i), using complexity information for streaming patients and urgency information for prioritizing them (complexity streaming) is better than using urgency information for streaming and complexity information for prioritizing them (urgency streaming). Furthermore, the performance advantage of complexity streaming (weakly) increases as total ED utilization increases.*

The intuition behind the above result is that matching capacity to workload in the different streams diminishes the effect of different service times among simple and complex patients. Hence, the difference in the intensity of adverse events between urgent and non-urgent patients becomes the dominant factor in selecting the best streaming design. Since in complexity streaming, patients in both streams are prioritized based on their urgency, it is more effective than urgency streaming. Although we have derived this insight using a single server model, we note that the single server assumption is not essential to the comparison. First, it can be easily seen from the well-known

Sakasegawa equation (see, e.g., Hopp and Spearman (2008) pp. 290–291) that with the same utilization, the queuing time of a $G/G/k$ queuing system and that of an “equivalent” $G/G/1$ become equal as the system’s utilization approaches 1. Since utilization in the ED is typically high, we expect the single server assumption to provide a good approximation. Second, our numerical comparisons show that even with utilization as low as 60%, the conclusion that complexity based streaming is superior to urgency based streaming is not altered by the number of servers. Finally, in Section 6.4, we use hospital data and a simulation model with multiple physicians, beds, etc. to further confirm the superiority of complexity streaming.

6. Simulation Analysis of Complexity-Augmented Triage

In this section, we test the conjectures suggested by our analytic models and get a better sense of the magnitude of the impact of complexity-augmented triage by means of a detailed ED simulation model. This simulation incorporates many features common to most ED’s, including dynamic non-stationary arrivals, multi-stage service, multiple physicians and exam rooms, inaccuracy in triage classifications (both in terms of urgency and complexity), and limits on the number of patients physicians handle simultaneously. We use a year of hospital data from the University of Michigan Hospital ED (UMHED) plus time study data from the literature to construct a base case that is representative of ED’s in research hospitals. We first describe the main features of our simulation framework, and then describe the test cases and our conclusions from them.

Patient Classes. At the time of triage, patients are classified according to both urgency (urgent or non-urgent) and complexity (simple or complex). For modeling purposes, we omit the resuscitation unit (RU) and fast track (FT) classifications, shown in Figure 1 (right), since these patients are typically tracked separately from the main ED. We define S (simple) patients as those who only require one interaction and C (complex) patients as those requiring two or more interactions. Note that we do not count the short physician visits at or after the disposition state as an interaction for the purpose of S/C classification. Focusing on the main area of the ED, with triage level 1, 4 and 5 patients omitted, we can equate U (N) patients with level 2 (level 3) patients. Since the majority of ED patients are composed of level 2 and 3 patients (about 80% in UMHED), improvements for this subset of patients will have a major impact on overall ED performance. Both urgency and complexity classifications at the point of triage are subject to errors with different error rates, so we assume the true type of a patient is not known until the final disposition decision is made. Consistent with the empirical findings of Hay et al. (2001), Vance and Spirvulis (2005), and Kronick and Desmond (2009), we assume urgency and complexity classifications are subject to 10% and 17% error rates, respectively. We also assume urgency-based and complexity-based misclassification rates are independent and symmetric (i.e., triage nurses are equally likely to classify U (C) patients as N (S) as they are to classify N (S) patients as U (C), respectively), but we consider asymmetric

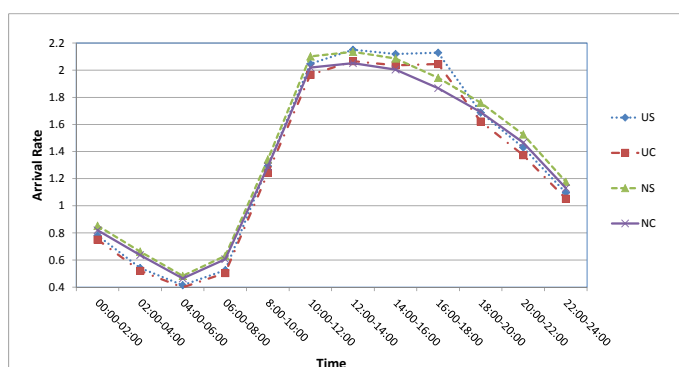


Figure 4 Class dependent arrival rates to the ED for an average day (obtained from a year of data in UMHED).

errors in our sensitivity analysis.

Arrival Process. Class-based patient arrivals are modeled using non-stationary Poisson processes that approximate our data. The non-stationary arrival rates for different classes are depicted in Figure 4. These arrival rates were obtained from a year of UMHED data taken at two-hour intervals. However, since UMHED patients are not currently triaged based on complexity, we assume that 49% of patients are complex as reported by Vance and Spirvulis (2005) in their large empirical study. The resulting pattern is similar to those reported in other studies (e.g., Green et al. (2006b)). A “thinning” mechanism (see Lewis and Shedler (1979a,b)) is used to simulate the non-stationary Poisson process arrivals for each class of patients. From our data and Figure 4, we also observe that complexity and urgency classifications are almost independent (e.g., a complex patient is equal likely to be urgent or non-urgent).

Service Process. The ED service process has multiple stages as depicted in Figure 3. Each patient experiences one or more patient-physician interactions followed by test/preparation/wait activities during which the physician cannot have a direct interaction with the patient (all such stages are labeled as Test in Figure 3). We also consider the initial and final preparations by a nurse. The initial preparation happens when the patient is moved to an exam room for the first time (before the first interaction with the physician) and the final preparation happens after the final interaction with the physician and before the patient is discharged home or admitted to the hospital. The duration of each physician interaction is random and can be approximated by an exponential distribution with a parameter that depends on the class of the patient as well as the number of previous interactions. Our data suggest that the first and last interactions are typically longer than the intermediate interactions, so we model them as such in the simulation. As mentioned above and illustrated in Figure 1 (right), S patients are defined as those requiring only one interaction. While S patients have 1 interaction, for C patients we simulate the distribution of the number of physician interactions using the data shown in Figure 5, which are derived from a detailed time study (see Table 3 of Graff et al. (1993)) with normalization to represent our NC and UC patient classes. The simulated service process is considered to be non-collaborative, since an ED physician rarely transfers his/her patients to another physician, and also non-preemptive.

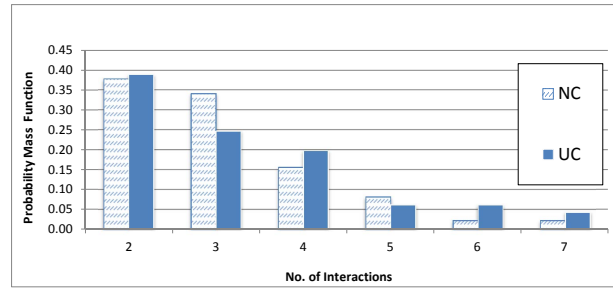


Figure 5 Probability mass function of the number of class-based interactions for complex patients (those requiring more than one interaction).

Physician-Patient Assignments and Priorities. As mentioned earlier, the process of connecting patients with physicians involves two phases. In Phase 1, patients are brought back from the waiting area to exam rooms whenever a room becomes available according to a Phase 1 sequencing rule. In Phase 2, whenever a physician becomes available, and if s/he has fewer than his/her maximum number of patients (7 is typical), s/he chooses the next patient from those available based on a Phase 2 sequencing rule, which can make use of information generated at triage. We assume that when urgency-based triage is used, U patients get priority over N patients in both Phases 1 and 2. If complexity-augmented triage is used, patients are prioritized in both Phases according to the priority ordering US, UC, NS, NC (ranked from high to low priority) which we found to be optimal in the simplified ED models discussed previously. When a patient is brought back to an examination room, we assume that s/he is assigned to the physician with the lowest number of patients. If all physicians are handling their limit of 7 patients, the patient must wait. Phase 1 and Phase 2 priority decisions can only be made based on the estimated class of the patient, which is subject to misclassification error, but adverse events are determined by the true class of the patient.

ED Resources. We consider 22 beds and 4 physicians in our base case scenario, which are representative of a medium sized ED. But we perform sensitivity analysis to understand the effect of number of both beds and physicians on the benefit of complexity-augmented triage. We also consider cases with non-stationary staffing in order to examine the effect of better matching staffing to the demand profile. We consider test facilities (ancillary services) as exogenous resources (i.e., test times are independent of the volume of ED patients) because these facilities typically handle many other patients besides those from the ED. Hence, tests and waiting for their reports result in various exogenous “delays,” which were approximated with class dependent exponential random variables and were estimated to have mean “delay” times roughly 2.5 times longer for complex patients than for simple patients based on UMHED data.

Adverse Events. Adverse events are simulated using Point processes with stationary rates that depend on patient class and phase of service. Specifically, in our base case we use class and phase dependent Poisson processes and assume that (i) U patients have a higher intensity of adverse events than N patients, and (ii) the intensity of adverse events decreases by 60% when patients

move from the waiting room (Phase 1) to an examination room (Phase 2). (The 60% number is a physician estimate based on the impact of more careful monitoring and care within the ED, but we have done sensitivity analysis that shows the main conclusions are robust to this estimate.) As in our previous models, we do not consider fatal events that would terminate the adverse events counting process, since the impact of these rare events on our objective function is extremely small.

Runs. The simulation was written in C++ and made use of a cyclo-stationary model (see, e.g., Gardner et al. (2006) for a complete review of cyclo-stationarity) with a period of a week. Each data point was obtained for 5000 replications of one week, where each replication was preceded by a warm-up period of one week. This was observed to be sufficient because correlations in the ED flow are very small for spans of two or more days due to the fact that ED's generally clear out overnight. The number of replications (5000) was chosen to achieve confidence intervals tight enough to (1) ensure the sample averages are reliable, and (2) allow omission of these very tight intervals from our data presentations.

In the following sections, we describe how we used our simulation model to analyze the benefit of complexity-augmented triage over urgency-based triage.

6.1. Performance of Complexity-Augmented Triage

We start by comparing complexity-based triage to urgency-based triage in our base case model, under the assumption that both types of triage make use of their respective priority rules for sequencing patients in both Phase 1 and Phase 2. This leads to the following:

Observation 1. *In the base case, implementing complexity-augmented triage rather than urgency-based triage improves ROAE and LOS by 9.4% (0.16 events/hr) and 7.6% (36 mins/patient), respectively.*

To consider the case where Phase 2 sequencing cannot follow the optimal rule due to a lack of data, patient discomfort, or other factors, we also compare complexity-augmented triage with urgency-based triage when Phase 2 sequencing in both systems uses a service-in-random-order (SIRO) rule. This leads to improvements of 7.9% and 7.0% in ROAE and LOS, respectively. Hence, it appears that the benefits of complexity-augmented triage are quite robust to the policy used in Phase 2. At least in our base case, it is the refined sequencing in Phase 1 that drives the majority of the improvement. Based on our sensitivity analyses, this conclusion is not significantly affected by many assumptions in the base case including the 60% drop in Phase 2 intensity of adverse events and the 2.5 ratio of test times of complex patients to simple patients.

The smaller effect of Phase 2 sequencing compared to that of Phase 1 prioritization is mainly due to the fact that, under the conditions of our base case, physicians in Phase 2 often do not have many available patients from which to choose. This is because (a) patients are unavailable for considerable amounts of time while being tested and waiting for test results, and (b) each physicians handles

only a limited number of patients simultaneously (with an upper bound of seven). However, in ED's with shorter test times (e.g., more test facilities dedicated to the ED, or more responsive central test facilities), larger case loads (patients per physician), and enough examination rooms/beds to accommodate patients, there will be more choices among in-process patients, and hence more improvement from an effective Phase 2 sequencing policy. To test this, we consider an ED with test rates 70% faster than the base case values, 40 beds, 3 physicians, and a maximum number of 10 patients per physician. Under these conditions, if Phase 2 sequencing is done according to SIRO for both the urgency-based and complexity-augmented triage systems, then complexity-augmented triage achieves improvements of 8.6% and 6.2% in ROAE and LOS, respectively, relative to urgency-based triage. In contrast, if the urgency-based triage system prioritizes patients in Phase 2 by urgency ($U > N$) and the complexity-augmented triage system prioritizes patients in Phase 2 by complexity and urgency ($US > UC > NS > NC$), then complexity-augmented triage achieves improvements of 13.1% and 9.10% in ROAE and LOS, respectively, relative to urgency-based triage. This leads us to the following:

Observation 2. *In ED's where physicians have more choice about what patient to see next, using complexity information to prioritize patients in Phase 2 becomes more valuable.*

All of the above results assume complex patients are those requiring at least two interactions. But we did perform sensitivity analysis on this choice, which confirmed that:

Observation 3. *If the number of interactions is used as the metric for patient complexity, the benefit of complexity-augmented triage is greatest when complex patients are defined to be those requiring at least two interactions.*

The reason for this is that increasing the number of interactions required for a patient to be considered complex decreases the fraction of complex patients substantially, but only slightly increases the difference in treatment times between complex and simple patients. Thus, as predicted by Proposition 3, the benefit of complexity-augmented triage declines.

6.2. The Effect of ED Resource Levels

Another factor predicted by Proposition 3 to favor complexity-augmented triage is resource utilization. In that proposition, resources refer to both physicians and examination rooms (which are indistinguishable in the single-stage simplified ED model). Hence, we expect higher utilization of either physicians or examination rooms to increase the benefit of complexity-augmented triage. Figure 6 illustrates the percentage improvement in terms of ROAE and LOS from using complexity-augmented triage over urgency-based triage for varying numbers of examination rooms and physicians. From this figure we observe the following:

Observation 4. *The benefit of complexity-augmented triage is greater in ED's with higher bed and/or physician utilization.*

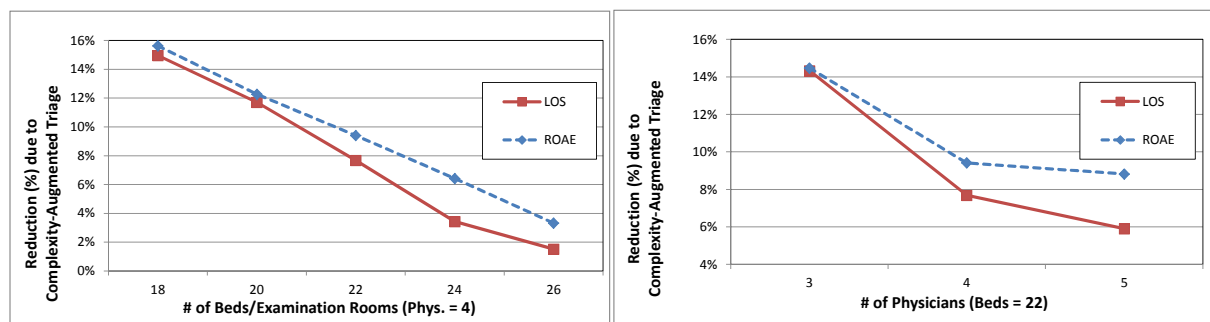


Figure 6 The effect of resources (beds and physicians) on the benefit of complexity-augmented triage over the current practice of urgency-based triage [Left: the effect of beds (4 physicians); Right: the effect of physicians (22 beds)].

As we observed in the Introduction, most ED's are overcrowded, so high utilization is commonplace. Hence, results from our analytic and simulation models suggest that complexity-augmented triage is most effective precisely in ED's most in need of improvement.

Non-Stationary Staffing. Because ED's typically adjust staffing to follow workload, at least to some extent, we now consider two cases of non-stationary staffing: (i) reducing the staffing level during off-peak hours, and (ii) reducing the staffing level during off-peak hours while increasing staffing during peak hours (redistributing the current workforce). To examine these cases, we consider two alternate scenarios that modify our base case assumption of 4 physicians at all times: (i) 4 physicians during peak demand times (12-hour shifts) and 3 physicians otherwise, and (ii) 6 physicians during peak times (12-hour shifts) and 2 otherwise, so there is no net change in labor hours. Under Scenario (i), the complexity-augmented triage achieves 11.5% and 11.0% improvements in ROAE and LOS, respectively, compared to urgency-based triage. Under Scenario (ii), these numbers are 8.8% and 6.9%. Hence, the improvements relative to the 9.4% and 7.4% improvements of the base case shown in Figure 6 (right) are larger under Scenario (i) but not under Scenario (ii). The reason is that Scenario (i) reduces staffing during off peak hours, which we have already shown enhances the benefits of complexity-augmented triage. But Scenario 2 increases utilization during peak hours, in addition to decreasing utilization during off peak hours. Since overall performance is dominated by the peak hours, during which most congestion occurs, this results in a net decrease in the benefits of complexity-augmented triage. Nevertheless, since it is not economical to entirely eliminate high utilization periods in the ED, the benefits of complexity-augmented triage will be reduced but not eliminated with staffing that better matches the demand profile.

Bed-Block Phenomenon. We can use the results of Figure 6 to predict the effect of the ED bed-block phenomenon, in which ED patients admitted to the hospital cannot be transferred to their inpatient unit due to unavailability of beds. By tying up beds in the ED to board admitted patients, bed-block reduces the effective number of ED beds, and hence, increases their utilization. Therefore, from Figure 6 and Observation 4, we can expect complexity-augmented triage to yield greater benefits in ED's with higher bed-block/boarding times. For more detailed discussion of the effect of ED bed-block on patient flow design, we refer interested readers to Saghaian et al. (2012)

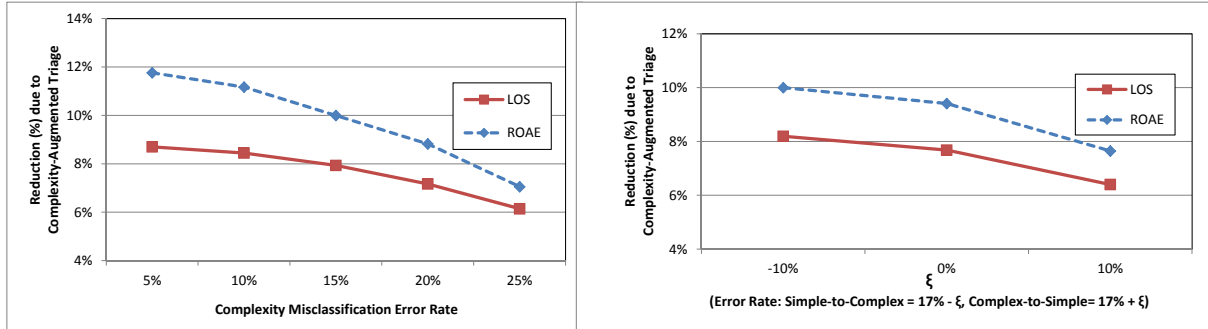


Figure 7 The effect of complexity misclassification error rates on the benefit of a complexity-augmented triage (compared to an urgency-based only) [Left: symmetric misclassification; Right: asymmetric misclassification].)

and the related references therein.

6.3. The Effect of Misclassification

Misclassification errors are inevitable in any triage system. Figure 7 (left) shows the improvement in ROAE and LOS achieved by complexity-augmented triage relative to urgency-based triage for variations of the base case, in which complexity misclassification error rates range from 5% to 25%. Figure 7 (left) assumes these errors to be symmetric; that is, the chance of classifying an S patient as C is equal to the chance of classifying a C patient as S. Figure 7 (right) considers asymmetric error rates while keeping the average misclassification rate constant and equal to the base-case value of 17%. From these figures, we observe the following:

Observation 5. *The benefit of complexity-augmented triage is relatively robust to complexity misclassification errors. However, complex-to-simple misclassifications are slightly more harmful than simple-to-complex misclassifications.*

The intuition behind the second part of this observation is that a complex-to-simple misclassification error moves a complex patient up in the queue, potentially delaying many other patients. In contrast, a simple-to-complex misclassification error moves a single simple patient back in the queue, delaying only that patient. So, it is slightly better to err on the side of classifying ambiguous patients as complex rather than simple.

6.4. Complexity Streaming Patient Flow Design

Finally, we return to the question of whether patient complexity information is most valuable in prioritizing or streaming patients. To do this, we examine a *complexity streaming* design in which patients are divided into two streams: one for patients triaged as simple (S) and one for those triaged as complex (C). The resources (beds and physicians) are labeled with S and C, indicating their main purpose. However, to overcome the “anti-pooling” disadvantage of streaming, we allow physicians or beds allocated to one stream to be used by the other stream in certain circumstances. When a C physician is available but there is no complex patient available, the physician can be assigned to an S patient who is waiting, and vice-versa. Also, an arrival of type S may enter a C bed if no other C patients are waiting, and vice-versa. This type of flexible allocation mechanism is

referred to as “virtual streaming” in Saghafian et al. (2012), which demonstrates it to be important in disposition based streaming protocols. Here we assume that patients in both streams and in both Phases 1 and 2 are prioritized according to their urgency level.

Separating simple and complex patients makes it easier to implement lean process improvement techniques to improve and standardize service, particularly on the simple side for which the repetitive treatment processes can be organized in a clear flow-shop manner (see also Clark and Huckman (2012) and KC and Terwiesch (2011) for related discussions on performance benefits of focused operations). Without separating simple patients from complex ones, lean process improvements are much more difficult to implement because many tasks will not be amenable to standardized procedures.

We first exclude the effect of lean improvements and compare the performance of *complexity streaming*, in which complexity information is used for streaming and urgency information is used for prioritizing, with *urgency streaming*, where urgency information is used for streaming and complexity information is used for prioritizing. We perform this comparison after optimizing the assignment of resources (physicians and beds) to each stream for each patient flow design. We observe that, even without lean improvements, using the complexity information for streaming and urgency information for prioritizing is better than using the urgency information for streaming and complexity information for prioritizing. This confirms our earlier result of Section 5.3 in a more realistic setting. We also compare, in Figure 8, the performance of complexity streaming, with and without lean improvements, against that of *urgency prioritization* (i.e., current practice in which patients in the main ED are not streamed but are prioritized based on urgency) and *complexity-augmented prioritization* (i.e., a design in which patients are not streamed but are prioritized in Phase 1 and Phase 2 according to the optimal priority rule using complexity-augmented triage information). The system with lean improvements assumes that these increase the service rate for interactions with simple patients by 10%, but that no change occurs for complex patients. Based on results in other industries, this is a conservative estimate of the impact of a lean transformation. Figure 8 compares performance in terms of LOS (results for the ROAE criterion are similar). These comparisons lead to the following observations:

Observation 6. *It is better to use complexity information for streaming and urgency information for prioritizing than using urgency information for streaming and complexity information for prioritizing (5.7% and 4.8% improvements in ROAE and LOS, respectively).*

Observation 7. *Without lean improvements, complexity streaming is better than the current practice (urgency prioritization), but worse than complexity-augmented prioritization. With lean improvements (made only to the simple stream), complexity streaming can achieve a substantial advantage over complexity-augmented prioritization.*

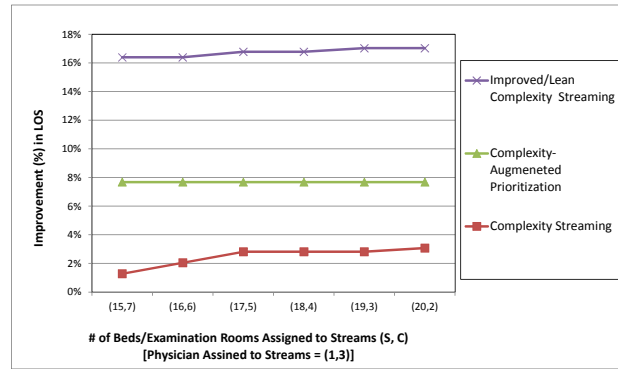


Figure 8 Performance of different patient flow designs compared to the current practice (i.e., urgency prioritization).

7. Conclusion

In this paper, we propose a new triage system for ED practice in which patients are classified on the basis of complexity, as well as urgency. Our results suggest that, compared to current urgency-based triage systems, complexity-augmented triage can significantly improve ED performance in terms of both patient safety (ROAE) and operational efficiency (LOS), even if patient classification is subject to error.

We show that complexity information gathered at triage can be used to improve patient prioritization in the ED. Our analyses indicate that the current practice of prioritizing patients purely based on urgency (e.g., triage level 2 over 3 in the main ED) is suboptimal compared to a system that takes into account a measure of patient complexity. Incorporating complexity can address many of the performance limitations of the current triage system in ED's that have been widely reported by clinicians (see, e.g., Welch and Davidson (2011) and the references therein).

To collect the needed complexity information, we find that a simple and fast classification scheme, which defines patients to be simple if they require only a single interaction (and complex otherwise) works very well as the basis for complexity-augmented triage because it results in (1) a nearly even split between simple and complex patients, and (2) a substantial difference between average treatment time of complex and simple patients. Empirical tests have shown that this classification scheme can be implemented at triage with reasonable accuracy.

Our analyses indicate that complexity-augmented triage can yield substantial safety and efficiency improvements even if complexity information is only used to prioritize patients up to entry into the examination rooms (Phase 1). Furthermore, in ED's where physicians have a significant amount of choice about what patient to see next within examination rooms (Phase 2), we find that complexity information gathered at triage can yield additional benefits by facilitating internal sequencing decisions. For both Phase 1 and Phase 2, the benefit of complexity-augmented triage is greatest in ED's with high physician and/or examination room utilization. Since ED's are widely overcrowded, our results suggest that complexity-augmented triage is an effective way for ED's to improve safety and reduce congestion without adding expensive human or physical capacity.

We investigate a new patient flow design, in which complexity-augmented triage information is used to separate simple and complex patients into two streams. Our results suggest that it is more effective to stream patients based on their complexity and then prioritize them within each based on their urgency than it is to stream them according to urgency and prioritize them according to complexity. Streaming based on complexity also facilitates implementation of lean methods in the “simple” patient stream, which takes advantage of complexity-augmented triage information to achieve even greater gains. If these gains are substantial enough, such complexity streaming can outperform a system in which complexity-information gathered at triage is used to augment patient prioritization without any streaming.

Our results indicate that even a simple estimate of patient complexity can be used to make significant improvements in safety and flow. Three future streams of research that could build on our insights to achieve even better performance are: (1) finding data driven rules, which correlate patient characteristics, symptoms and evaluations to treatment time and resource requirements, and can serve as the basis of even more effective prioritization and streaming policies than those suggested here; (2) developing statistical tools for tracking patient class dependent delays in test facilities and analytic models for incorporating these into Phase 1 and Phase 2 sequencing rules, and (3) constructing dynamic patient prioritization systems that make use of real-time information on patient and resource status to sequence patients into and through the ED. All of these enhancements could be used in the context of a single ED or within streams set up to facilitate standardization efficiencies. Whether and how the performance improvements from these systems can justify their implementational complexity relative to the simple system we have proposed here is an open research question.

References

- Argon, N.T., S. Ziya. 2009. Priority assignment under imperfect information on customer type identities. *Manufacturing Service Oper. Management* **11**(4) 674–693.
- Baker, M., M. Clancy. 2006. Can mortality rates for patients who die within the emergency department, within 30 days of discharge from the emergency department, or within 30 days of admission from the emergency department be easily measured? *Emerg. Med. J.* **23**(8) 601–603.
- Batt, R.J., C. Terwiesch. 2012. Doctors under load: an empirical study of state-dependent service times in emergency care. Working Paper, Wharton School of Business.
- Ben-Tovim, D. I., J. E. Bassham, D. M. Bennett, M. L. Dougherty et al. 2008. Redesigning care at the Flinders Medical Centre: clinical process redesign using lean thinking. *Medical Journal of Australia* **188**(6) 27–31.
- Brennan, T.A., L.L. Leape, M.N. Laird et al. 1991. Incidence of adverse events and negligence in hospitalized patients. *The New England J. of Med.* **324**(6) 370–376.
- Buyukkoc, C., P. Varaiya, J. Walrand. 1985. The $c\mu$ rule revisited. *Adv. Appl. Prob.* **17** 237–238.
- Clark, J.R., R.S. Huckman. 2012. Broadening focus: Spillovers, complementarities, and specialization in the hospital industry. *Management Sci.* **58**(4) 708–722.

- Cobham, A. 1954. Priority assignment in waiting line problems. *J. Oper. Res. Soc. of Amer.* **2** 70–76.
- Cobham, A. 1955. Priority assignment - a correction. *J. Oper. Res. Soc. of Amer.* **3** 547.
- Cox, D.R., W.L. Smith. 1961. *Queues*. Methuen & Co, London.
- Diercks, D.B., M.T. Roe, A.Y. Chen, W.F. Peacock et al. 2007. Prolonged Emergency Department stays of nonsegment- elevation myocardial infarction patients are associated with worse adherence to the American College of Cardiology/American Heart Association guidelines for management and increased adverse events. *Annals of Emerg. Med.* **50**(5) 489–496.
- Dobson, G., T. Tezcan, V. Tilson. 2013. Optimal workflow decisions for investigators in systems with interruptions. *Management Sci.* **59**(5) 1125–1141.
- Fernandes, C.M., P. Tanabe, N. Gilboy et al. 2005. Five level triage: a report from the ACEP/ ENA five-level task force. *J. Emerg. Nurs.* **31** 39–50.
- FitzGerald, G., G.A. Jelinek, D. Scott, M.F. Gerdtz. 2010. Emergency Department triage revisited. *Emergency Medicine Journal* **27** 86–92.
- Gardner, W. A., A. Napolitano, L.Paura. 2006. Cyclostationarity: Half a century of research. *Signal Processing* **86** 639–697.
- Gilboy, N., P. Tanabe, D.A. Travers, A.M. Rosenau, D.R. Eitel. 2005. *Emergency Severity Index, Version 4: Implementation Handbook*. Agency for Healthcare Research and Quality Publication No. 05-0046-2, Rockville, MD.
- Graff, L. G., S. Wolf, R. Dinwoodie, D. Buono, D. Mucci. 1993. Emergency physician workload: A time study. *Annals of Emergency Medicine* **22**(7) 1156–1163.
- Green, L., S. Savin, B. Wang. 2006a. Managing patient service in a diagnostic medical facility. *Oper. Res.* **54**(1) 11–25.
- Green, L. V., J. Soares, J. F. Giglio, R.A. Green. 2006b. Using queuing theory to increase the effectiveness of Emergency Department provider staffing. *Academic Emergency Medicine* **13**(1) 61–68.
- Guo, X., K. Liu. 2001. A note on optimality conditions for continuous-time Markov Decision Processes with average cost criterion. *IEEE Trans. on Aut. Contr.* **46**(12) 1984–1989.
- Hay, E., L. Bekerman, G. Rosenberg, R. Peled. 2001. Quality assurance of nurse triage: Consistency of results over three years. *American J. of Emerg. Med.* **19**(2) 113–117.
- Hopp, W.J., M.L. Spearman. 2008. *Factory Physics*. Third Edition. McGraw-Hill, Burr Ridge, IL.
- Huang, J., B. Carmeli, A. Mandelbaum. 2012. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. Working Paper, National University of Singapore.
- Ierson, K.V., J.C. Moskop. 2007. Triage in medicine, part I: Concept, history, and types. *Annals of Emerg. Med.* **49**(3) 275–281.
- Kakalik, J.S., J.D.C. Little. 1971. *Optimal Service Policy for the M/G/1 Queue with Multiple Classes of Arrival*. Rand Corporation Report.
- KC, D.S., C. Terwiesch. 2011. The effects of focus on performance: Evidence from California hospitals. *Management Sci.* **57**(19) 1897–1912.
- Keen, W.W. 1917. *The Treatment of War Wounds*. W.B. Saunders, Philadelphia, PA.
- King, D. L., D. I. Ben-Tovim, J. Bassham. 2006. Redesigning Emergency Department patient flows: Application of lean thinking to health care. *Emergency Medicine Australasia* **18** 391–397.
- Kronick, S.L., J.S. Desmond. 2009. Blink: Accuracy of physician estimates of patient disposition at the time of ED triage. SAEM Midwest Regional Meeting.
- Lewis, P. A.W., G. S. Shedler. 1979a. Simulation of nonhomogenous Poisson processes by thinning. *Naval Research Logistics Quarterly* **26**(3) 403–413.

- Lewis, P. A.W., G. S. Shedler. 1979b. Simulation of nonhomogenous Poisson processes with degree-two exponential polynomial rate function. *Oper. Res.* **27**(5) 1026–1039.
- Lippman, S. 1975. Applying a new device in the optimization of exponential queueing system. *Oper. Res.* **23**(4) 687–710.
- Liu, S.W., S.H. Thomas, J.A. Gordon, J. Weissman. 2005. Frequency of adverse events and errors among patients boarding in the emergency department. *Acad. Emerg. Med.* **12** 49b–50b.
- Mills, A.F., N.T. Argon, S. Ziya. 2013. Resource-based patient prioritization in mass-casualty incidents. *Manufacturing Service Oper. Management* **15**(3) 361–377.
- Moskop, J.C., K.V. Ierson. 2007. Triage in medicine, part II: Underlying values and principles. *Annals of Emerg. Med.* **49**(3) 282–287.
- Patrick, J., M.L. Putterman, M. Queyranne. 2008. Dynamic multipriority patient scheduling for a diagnostic resource. *Oper. Res.* **56**(6) 1507–1525.
- Saghafian, S., W.J. Hopp, M.P. Van Oyen, J.S. Desmond, S.L. Kronick. 2012. Patient streaming as a mechanism for improving responsiveness in Emergency Departments. *Oper. Res.* **60**(5) 1080–1097.
- Saghafian, S., M.P. Van Oyen, B. Kolfal. 2011. The “W” network and the dynamic control of unreliable flexible servers. *IIE Transactions* **43**(12) 893–907.
- Scheuermeyer, F.X., J. Christenson, G. Innes et al. 2010. Safety of assessment of patients with potential ischemic chest pain in an emergency department waiting room: A prospective comparative cohort study. *Annals of Emergency Medicine* **56**(5) 455–462.
- Shaked, M., J.G. Shanthikumar. 2007. *Stochastic Orders*. Springer, New York, NY.
- Siddharathan, K., W.J. Jones, J.A. Johnson. 1996. A priority queueing model to reduce waiting times in emergency care. *International J. of Health Care Quality Assurance* **9**(5) 10–16.
- Tcha, D.W., S.R. Pliska. 1977. Optimal control of single-server queueing networks and multi-class M/G/1 queues with feedback. *Oper. Res.* **27**(2) 248–258.
- van der Zee, S.P., H. Theil. 1961. Priority assignment in waiting-line problems under conditions of misclassification. *Oper. Res.* **9** 875–885.
- Van Mieghem, J. A. 1995. Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *Annals of Appl. Prob.* **5**(3) 809–833.
- Vance, J., P. Spirvulis. 2005. Triage nurses validly and reliably estimate Emergency Department patient complexity. *Emergency Medicine Australasia* **17** 382–386.
- Wang, Q. 2004. Modeling and analysis of high risk patient queues. *Eur. J. of Oper. Res.* **155** 502–515.
- Welch, S.J., S.J. Davidson. 2011. The performance limits of traditional triage. *Annals of Emerg. Med.* **58**(2) 143–144.

Online Appendix A (Proofs)

Proof of Theorem 1: To show the result, we use an *interchange* argument; we show that if classes $uc \in \mathcal{U} \times \mathcal{C}$ and $sl \in \mathcal{U} \times \mathcal{C}$ are such that $p'_{uc} \hat{\theta}'_{uc} \hat{\mu}'_{uc} \geq p'_{sl} \hat{\theta}'_{sl} \hat{\mu}'_{sl}$, then it is (weakly) better to serve class uc than class sl when in state $(\underline{x}, \underline{y})$ with $x_{uc}, x_{sl} > 0$. This will also prove that the optimal policy will not idle the physician when there are one or more patients available in the rooms, since idling can be thought of serving an additional class, class 0, with $\hat{\theta}'_0 = \hat{\mu}'_0 = p'_0 = 0$ (see, for instance, Buyukkoc et al. (1985)). To show that it is (weakly) better to serve class uc than class sl , we first consider the problem in an N -period discounted cost setting with four parallel (one for each class of patients) $\cdot/M/k$ systems (to guarantee bounded transition rates for the purpose of uniformization) in place of the $\cdot/M/\infty$ test stage, and show that the results hold for any number of periods to go $n \in 1, 2, \dots, N$. (Notice that using four parallel $\cdot/M/k$ systems removes the need for considering the sequence and the type of patients within the common queue.) Using a convergence argument, as $n \rightarrow \infty$, it then follows that the result is true for an infinite-horizon (and hence, average cost) scenario with the four k -server test system. Next, taking limit as $k \rightarrow \infty$, it follows that the result is true even when transition rates are not bounded due to the existence of the $\cdot/M/\infty$ stage.

Now consider the finite horizon discounted cost version of (1). With β denoting the discount factor, the optimal discounted cost when there are $n + 1$ (uniformized) periods to go is

$$V_{n+1}^k(\underline{x}, \underline{y}) = \frac{1}{\psi_k} \left[\hat{\theta}'(\underline{x} + \underline{y})^T + \beta \left[\sum_{ij \in \mathcal{U} \times \mathcal{C}} [\lambda'_{ij} V_n^k(\underline{x} + \underline{e}_{ij}, \underline{y}) + (y_{ij} \wedge k) \eta V_n^k(\underline{x} + \underline{e}_{ij}, \underline{y} - \underline{e}_{ij})] \right. \right. \\ \left. \left. + \min_{a \in \mathcal{A}(\underline{x})} \left\{ \sum_{ij \in \mathcal{U} \times \mathcal{C}} \mathbb{1}_{\{a=ij\}} \hat{\mu}'_{ij} [p'_{ij} V_n^k(\underline{x} - \underline{e}_{ij}, \underline{y}) + (1 - p'_{ij}) V_n^k(\underline{x} - \underline{e}_{ij}, \underline{y} + \underline{e}_{ij})] \right. \right. \right. \\ \left. \left. \left. + (\psi_k - \sum_{ij \in \mathcal{U} \times \mathcal{C}} [\lambda'_{ij} + (y_{ij} \wedge k) \eta + \mathbb{1}_{\{a=ij\}} \hat{\mu}'_{ij}]) V_n^k(\underline{x}, \underline{y}) \right\} \right] \right], \quad (8)$$

or equivalently (grouping the terms related to control in the minimization and self-loop)

$$V_{n+1}^k(\underline{x}, \underline{y}) = \frac{1}{\psi_k} \left[\hat{\theta}'(\underline{x} + \underline{y})^T + \beta \left[\sum_{ij \in \mathcal{U} \times \mathcal{C}} [\lambda'_{ij} V_n^k(\underline{x} + \underline{e}_{ij}, \underline{y}) + (y_{ij} \wedge k) \eta V_n^k(\underline{x} + \underline{e}_{ij}, \underline{y} - \underline{e}_{ij})] \right. \right. \\ \left. \left. - \max_{a \in \mathcal{A}(\underline{x})} \left\{ \sum_{ij \in \mathcal{U} \times \mathcal{C}} \mathbb{1}_{\{a=ij\}} \hat{\mu}'_{ij} [p'_{ij} \Delta_{ij}^y V_n^k(\underline{x} - \underline{e}_{ij}, \underline{y}) + \Delta_{ij}^{x,y} V_n^k(\underline{x} - \underline{e}_{ij}, \underline{y} + \underline{e}_{ij})] \right. \right. \right. \\ \left. \left. \left. + (\psi_k - \sum_{ij \in \mathcal{U} \times \mathcal{C}} [\lambda'_{ij} + (y_{ij} \wedge k) \eta]) V_n^k(\underline{x}, \underline{y}) \right\} \right] \right], \quad (9)$$

where $\Delta_{ij}^y V_n^k(\underline{x}, \underline{y}) = V_n^k(\underline{x}, \underline{y} + \underline{e}_{ij}) - V_n^k(\underline{x}, \underline{y})$ and $\Delta_{ij}^{x,y} V_n^k(\underline{x}, \underline{y}) = V_n^k(\underline{x} + \underline{e}_{ij}, \underline{y} - \underline{e}_{ij}) - V_n^k(\underline{x}, \underline{y})$. Now let π ($\hat{\pi}$) be the policy that prescribes serving patients of class uc (sl) for every state $(\underline{x}, \underline{y})$ with $x_{uc}, x_{sl} > 0$ and in every period n . From (9), to show that π is (weakly) better than $\hat{\pi}$ in every period, we need to show that the following property holds for every n and every state $(\underline{x}, \underline{y})$ with $x_{uc}, x_{sl} > 0$:

$$\hat{\mu}'_{uc} [p'_{uc} \Delta_{uc}^y V_n^{k,\pi}(\underline{x} - \underline{e}_{uc}, \underline{y}) + \Delta_{uc}^{x,y} V_n^{k,\pi}(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc})]$$

$$\geq \hat{\mu}'_{sl} [p'_{sl} \Delta_{sl}^y V_n^{k,\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y}) + \Delta_{sl}^{x,y} V_n^{k,\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y} + \underline{e}_{sl})]. \quad (10)$$

To show property (10), we use induction on n . First, for $n = 0$, the property trivially holds since $V_0^\pi(\cdot, \cdot) = V_0^{\hat{\pi}}(\cdot, \cdot) = 0$. Next, suppose the property holds for n . We show that it will then also hold for $n + 1$. To do so, we need to consider different cases based on the state (i.e., partitions of the state space). First, consider the case where $x_{uc}, x_{sl} \geq 2$. Using action $a = uc$ (policy π) in both states $(\underline{x} - \underline{e}_{uc}, \underline{y})$ and $(\underline{x}, \underline{y})$ to compute $V_{n+1}^{k,\pi}(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc})$ and $V_{n+1}^{k,\pi}(\underline{x} - \underline{e}_{uc}, \underline{y})$ using (9), and subtracting the results we have $\Delta_{uc}^y V_{n+1}^{k,\pi}(\underline{x} - \underline{e}_{uc}, \underline{y}) =$

$$\begin{aligned} & \frac{1}{\psi_k} \left[\hat{\theta}'_{uc} + \beta \left[\sum_{ij \in \mathbf{U} \times \mathbf{C}} [\lambda'_{ij} \Delta_{uc}^y V_n^{k,\pi}(\underline{x}, \underline{y}) + (y_{ij} \wedge k) \eta \Delta_{uc}^y V_n^{k,\pi}(\underline{x} + \underline{e}_{ij} - \underline{e}_{uc}, \underline{y} - \underline{e}_{ij})] \right. \right. \\ & \quad \left. \left. + \mathbb{1}_{\{y_{uc} < k\}} \eta V_n^{k,\pi}(\underline{x}, \underline{y} - \underline{e}_{uc}) \right. \right. \\ & \quad \left. \left. - \hat{\mu}'_{uc} [p'_{uc} \Delta_{uc}^y \Delta_{uc}^y V_n^{k,\pi}(\underline{x} - 2\underline{e}_{uc}, \underline{y}) + \Delta_{uc}^y \Delta_{uc}^{x,y} V_n^{k,\pi}(\underline{x} - 2\underline{e}_{uc}, \underline{y} + \underline{e}_{ij})] \right. \right. \\ & \quad \left. \left. + (\psi_k - \sum_{ij \in \mathbf{U} \times \mathbf{C}} [\lambda'_{ij} + (y_{ij} \wedge k) \eta]) \Delta_{uc}^y V_n^{k,\pi}(\underline{x} - \underline{e}_{uc}, \underline{y}) - \mathbb{1}_{\{y_{uc} < k\}} \eta V_n^{k,\pi}(\underline{x} - \underline{e}_{uc}, \underline{y}) \right] \right]. \quad (11) \end{aligned}$$

Similarly, we can derive $\Delta_{uc}^{x,y} V_{n+1}^{k,\pi}(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc})$ using (9) and action $a = uc$ (policy π) in both states $(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc})$ and $(\underline{x}, \underline{y})$ and subtracting the results. Doing so we have $\Delta_{uc}^{x,y} V_{n+1}^{k,\pi}(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc}) =$

$$\begin{aligned} & \frac{1}{\psi_k} \left[\beta \left[\sum_{ij \in \mathbf{U} \times \mathbf{C}} [\lambda'_{ij} \Delta_{uc}^{x,y} V_n^{k,\pi}(\underline{x}, \underline{y} + \underline{e}_{uc}) + (y_{ij}^+ \wedge k) \eta \Delta_{uc}^{x,y} V_n^{k,\pi}(\underline{x} + \underline{e}_{ij} - \underline{e}_{uc}, \underline{y} - \underline{e}_{ij} + \underline{e}_{uc})] \right. \right. \\ & \quad \left. \left. - \mathbb{1}_{\{y_{uc} < k\}} \eta V_n^{k,\pi}(\underline{x}, \underline{y}) \right. \right. \\ & \quad \left. \left. - \hat{\mu}'_{uc} [p'_{uc} \Delta_{uc}^{x,y} \Delta_{uc}^y V_n^{k,\pi}(\underline{x} - 2\underline{e}_{uc}, \underline{y} + \underline{e}_{uc}) + \Delta_{uc}^{x,y} \Delta_{uc}^y V_n^{k,\pi}(\underline{x} - 2\underline{e}_{uc}, \underline{y} + \underline{e}_{ij} + \underline{e}_{uc})] \right. \right. \\ & \quad \left. \left. + (\psi_k - \sum_{ij \in \mathbf{U} \times \mathbf{C}} [\lambda'_{ij} + (y_{ij}^+ \wedge k) \eta]) \Delta_{uc}^{x,y} V_n^{k,\pi}(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc}) + \mathbb{1}_{\{y_{uc} < k\}} \eta V_n^{k,\pi}(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc}) \right] \right], \quad (12) \end{aligned}$$

where $y_{ij}^+ = y_{ij}$ for all $ij \neq uc \in \mathbf{U} \times \mathbf{C}$, and $y_{uc}^+ = y_{uc} + 1$. In a similar way and by using action $a = sl$ (policy $\hat{\pi}$) in (9), quantities $\Delta_{sl}^y V_{n+1}^{k,\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y})$ and $\Delta_{sl}^{x,y} V_{n+1}^{k,\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y} + \underline{e}_{sl})$ can be computed. Next, to check property (10) for $n + 1$, multiply (11) by $p'_{uc} \hat{\mu}'_{uc}$, and (12) by $\hat{\mu}'_{uc}$ and add up the results. Similarly, multiply $\Delta_{sl}^y V_{n+1}^{k,\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y})$ and $\Delta_{sl}^{x,y} V_{n+1}^{k,\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y} + \underline{e}_{sl})$ by $p'_{sl} \hat{\mu}'_{sl}$ and $\hat{\mu}'_{sl}$, respectively, and add up the results. Next, using the induction hypothesis and that $p'_{uc} \hat{\theta}'_{uc} \hat{\mu}'_{uc} \geq p'_{sl} \hat{\theta}'_{sl} \hat{\mu}'_{sl}$, after algebraic simplification it follows that

$$\begin{aligned} & \hat{\mu}'_{uc} [p'_{uc} \Delta_{uc}^y V_{n+1}^{k,\pi}(\underline{x} - \underline{e}_{uc}, \underline{y}) + \Delta_{uc}^{x,y} V_{n+1}^{k,\pi}(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc})] \\ & \quad - \hat{\mu}'_{sl} [p'_{sl} \Delta_{sl}^y V_{n+1}^{k,\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y}) + \Delta_{sl}^{x,y} V_{n+1}^{k,\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y} + \underline{e}_{sl})] \geq 0, \quad (13) \end{aligned}$$

which establishes property (10) for $n + 1$ for the case where $x_{uc}, x_{sl} \geq 2$. In a similar way, this property can be established for other cases (i.e., the remaining partition of the state space). Hence, the non-idling strict priority rule is optimal for all n . Next, taking the limit as $n \rightarrow \infty$ it follows

that the finite horizon problem converges to the infinite horizon one both in policy and cost (see Sennott (1999) Proposition 4.3.1). Furthermore, the convergence of the policy of the infinite-horizon discounted cost problem to that of average cost can easily be established (see Sennott (1999) Corollary 7.5.10). Therefore, the underlying non-idling strict priority policy is optimal under the average cost setting indexed by k (i.e., with $\cdot/M/k$'s in place of the $\cdot/M/\infty$) for any finite k . Since the result is true for any k , a convergence argument can be used to show that the result holds for the original problem with $k = \infty$. Notice that the existence of an optimal stationary policy for the original CTMS (i.e., when $k = \infty$) follows from the results of Guo and Liu (2001). \square

Proof of Proposition 1: The proof of part (i) follows directly from comparing (5) and (6). To show part (ii), notice that, using the result of part (i) for a special case where there is no misclassification error, prioritizing U (N) patients is optimal if, and only if, $\theta_U \mu_U \geq (\leq) \theta_N \mu_N$. Next, observe that $\theta'_U \mu'_U - \theta'_N \mu'_N = [\lambda_N \lambda_U \mu_N \mu_U (\theta_U \mu_U - \theta_N \mu_N) (1 - \gamma_N - \gamma_U)] / [(\lambda_N \mu_U \gamma_N + \lambda_U \mu_N (1 - \gamma_U)) (\lambda_N \mu_U (1 - \gamma_N) + \lambda_U \mu_N \gamma_U)]$. Combining these two results completes the proof of part (ii), as the sign of the numerator changes when the sum of errors exceeds 1. \square

LEMMA 1 (Perfect Classification - Prioritization). *In the simplified single-stage ED model under perfect urgency and complexity based classification:*

(i) *The best priority rule is to prioritize patients in decreasing order of $\theta\mu$ values. Hence, if $\theta_{UC}\mu_{UC} \geq \theta_{NS}\mu_{NS}$, then the best priority rule is to follow the ordering: US, UC, NS, NC. Otherwise, the ED should follow the priority ordering: US, NS, UC, NC.*

(ii) *$R_*^{\mathcal{U}\cup\mathcal{C}} \leq R_*^{\mathcal{U}}$. That is, the risk of adverse events under the optimal priority rule using both complexity and urgency information is (weakly) smaller than that under the optimal apriority rule using only urgency information.*

(iii) *The best priority rule of part (i) is optimal even among the larger class of all non-anticipative policies (state or history dependent, idling or non-idling, etc.).*

Proof of Lemma 1: Notice that, using (2), we can compute the average waiting time of each class of patients under any (static) priority rule. Furthermore, under priority rule π , we have

$$R_\pi^{\mathcal{U}\cup\mathcal{C}} = \sum_{i \in \mathcal{U}} \sum_{j \in \mathcal{C}} \theta_{ij} \lambda_{ij} W_{ij}^\pi, \quad (14)$$

where W_{ij}^π is the average waiting of class ij under priority rule π . The proof of part (i) then follows from Cox and Smith (1961) (see pages 83-84), where an interchange argument is used (when the number of customer classes is at least 3) to show that the best rule (among the priority policies) to minimize the holding cost in a non-preemptive M/G/1 is to follow the $c\mu$ rule. Replacing holding cost values (c) with adverse event rates (θ), and noticing that the patient class US (NC) has the highest (lowest) $\theta\mu$ value complete the proof of part (i). Next, using the result of part (i) together with (2) and (14), when $\theta_{UC}\mu_{UC} \geq \theta_{NS}\mu_{NS}$, we have:

$$\begin{aligned}
R_*^{\mathcal{U}\cup\mathcal{C}} &= \lambda \mathbb{E}(s^2) \left[\frac{\lambda_{US}\theta_{US}}{2(1-\rho_{US})} + \frac{\lambda_{UC}\theta_{UC}}{2(1-\rho_{US})(1-\rho_{US}-\rho_{UC})} + \frac{\lambda_{NS}\theta_{NS}}{2(1-\rho_{US}-\rho_{UC})(1-\rho_{US}-\rho_{UC}-\rho_{NS})} \right. \\
&\quad \left. + \frac{\lambda_{NC}\theta_{NC}}{2(1-\rho_{US}-\rho_{UC}-\rho_{NS})(1-\rho_{US}-\rho_{UC}-\rho_{NS}-\rho_{NC})} \right] \\
&\leq \min\{R_U^{\mathcal{U}}, R_N^{\mathcal{U}}\} = R_*^{\mathcal{U}},
\end{aligned} \tag{15}$$

where the inequality follows from (3) and (4) together with the result of part (i) of Proposition 1 (for the special case where there is no misclassification error).

When $\theta_{UC}\mu_{UC} < \theta_{NS}\mu_{NS}$, we have:

$$\begin{aligned}
R_*^{\mathcal{U}\cup\mathcal{C}} &= \lambda \mathbb{E}(s^2) \left[\frac{\lambda_{US}\theta_{US}}{2(1-\rho_{US})} + \frac{\lambda_{NS}\theta_{NS}}{2(1-\rho_{US})(1-\rho_{US}-\rho_{NS})} + \frac{\lambda_{UC}\theta_{UC}}{2(1-\rho_{US}-\rho_{NS})(1-\rho_{US}-\rho_{NS}-\rho_{UC})} \right. \\
&\quad \left. + \frac{\lambda_{NC}\theta_{NC}}{2(1-\rho_{US}-\rho_{UC}-\rho_{NS})(1-\rho_{US}-\rho_{UC}-\rho_{NS}-\rho_{NC})} \right],
\end{aligned} \tag{16}$$

and similar to the previous case, it can be easily seen that $R_*^{\mathcal{U}\cup\mathcal{C}} \leq R_*^{\mathcal{U}}$. The proof of part (iii) follows from Kakalik and Little (1971) (after replacing holding cost with adverse event rates) who (for the average holding cost objective) showed that the $c\mu$ policy of Cox and Smith (1961) remains optimal even when inserting idleness is allowed and/or when the priority rule is dynamic (i.e., state-dependent). \square

Proof of Proposition 2: The proof of part (i) follows directly from the proof of part (i) of Lemma 1, since all rates are replaced with their error impacted counterparts. That is, the same interchange method of Cox and Smith (1961) (see pages 83-84) after replacing all rates with their error impacted counterparts proves that the best priority rule is to give priority based on a decreasing order of $\theta\mu$ values. The proof of part (ii) follows from the proof of Lemma 1 (found earlier in this appendix) part (ii) after replacing parameters with their error impacted counter parts. The proof of part (iii) follows from the result of Kakalik and Little (1971), after replacing holding cost with the error impacted intensity of adverse events, and all the other rates with their error impacted counterparts. \square

LEMMA 2 (Perfect Classification - Attractiveness). *In the simplified single-stage ED model, perfect complexity-augmented triage yields a larger improvement over perfect urgency-based triage when (i) ED utilization is higher, (ii) heterogeneity in the average service time of simple vs. complex patients is larger, and/or (iii) the fraction of simple and complex patients are closer to equal.*

Proof of Lemma 2: To show the result, first consider the case where under the $\mathcal{U}\cup\mathcal{C}$ classification it is optimal to follow the priority order US, UC, NS, NC, and under the \mathcal{U} classification, it is optimal to follow the priority order U, N (i.e., prioritizing urgent patients first). Let $f = R_*^{\mathcal{U}\cup\mathcal{C}} - R_*^{\mathcal{U}}$, and notice that with $\mu_{i\mathcal{C}} = \mu_{\mathcal{C}}$ and $\mu_{i\mathcal{S}} = \mu_{\mathcal{S}}$ ($\forall i \in \mathcal{U}$), and $\theta_{Uj} = \theta_U$ and $\theta_{Nj} = \theta_N$ ($\forall j \in \mathcal{C}$) (i.e., when complexity is based only on set \mathcal{C} and urgency is based only on set \mathcal{U}), from (15) and (3) we have:

$$f = -\left[\frac{\theta_U \lambda_{US} \lambda_{UC} (1/\mu_C - 1/\mu_S)}{2(1 - \rho_U)} + \frac{\theta_N \lambda_{NC} \lambda_{NS} (1/\mu_C - 1/\mu_S)}{2(1 - \rho_U)(1 - \rho)}\right]. \quad (17)$$

Then, a careful treatment of utilization (realizing that $\rho_U = \lambda_U/\mu_U$ and $\rho = \rho_U + \rho_N$) shows that f is non-increasing in utilization, ρ . To prove part (ii), it then can be seen that f is non-increasing in $1/\mu_C - 1/\mu_S$ (keeping utilization and other factors the same). To see part (iii), let $\alpha \in [0, 1]$ denote the fraction of patients that are complex, and $(1 - \alpha)$ denote the fraction of patients that are simple, so $\lambda_{US} = (1 - \alpha)\lambda_U$, $\lambda_{UC} = \alpha\lambda_U$, $\lambda_{NC} = \alpha\lambda_N$, and $\lambda_{NS} = (1 - \alpha)\lambda_N$. Replacing these in (17), it follows that f , as a function of α , can be written as $f = -[\alpha(1 - \alpha)]k$, for some constant $k \geq 0$. Thus, $\alpha = 0.5$ yields the maximum benefit. The proof for other cases (i.e., when other priority rules are optimal) follows a similar argument after computing f using either (15) or (16), and either (3) or (4), depending on the optimal priority rule under $\mathbf{U} \cup \mathbf{C}$ and \mathbf{U} classifications, respectively. \square

Proof of Proposition 3: The proof of parts (i) - (iii) follows mainly from the proof of Lemma 2. First, consider the case where under the $\mathbf{U}' \cup \mathbf{C}'$ (i.e., imperfect urgency and complexity) classification it is optimal to follow the priority order US, UC, NS, NC, and under the \mathbf{U}' (i.e., imperfect urgency) classification, it is optimal to follow the priority order U, N (i.e., prioritizing urgent patients over non-urgent patients). With $f = R_{*}^{\mathbf{U}' \cup \mathbf{C}'} - R_{*}^{\mathbf{U}'}$, and after replacing rates with their error impacted counterparts in (17) we have:

$$f = -\left[\frac{\theta'_U \lambda'_{US} \lambda'_{UC} (1/\mu'_C - 1/\mu'_S)}{2(1 - \rho'_U)} + \frac{\theta'_N \lambda'_{NC} \lambda'_{NS} (1/\mu'_C - 1/\mu'_S)}{2(1 - \rho'_U)(1 - \rho')}\right]. \quad (18)$$

Next, notice that $\rho' = \rho$ (i.e., the total utilizations with and without misclassifications are the same). Hence, similar to the proof of part (i) of Lemma 2, it can be seen that f is non-increasing in ρ . Moreover, it can be seen that f is non-increasing in $1/\mu'_C - 1/\mu'_S$. Next, notice that $(1/\mu')^T = (A(\underline{\lambda}/\underline{\mu})^T)/\underline{\lambda}'$, where A is defined in (7). Thus, similar to the proof of part (ii) of Lemma 2, it can be seen that f is non-increasing in $1/\mu_C - 1/\mu_S$, which proves part (ii). Furthermore, similarly to the proof of part (iii) of Lemma 2, let $\lambda_{US} = (1 - \alpha)\lambda_U$, $\lambda_{UC} = \alpha\lambda_U$, $\lambda_{NC} = \alpha\lambda_N$, and $\lambda_{NS} = (1 - \alpha)\lambda_N$. It can be seen that f as a function of α is minimized at $\alpha = 0.5$, which proves part (iii). It can also be seen that f is non-decreasing in complexity misclassification error rates, γ_S and γ_C , which proves part (iv). The proof for other cases (i.e., when other priority rules are optimal) follows a similar line of argument after computing f . \square

Proof of Proposition 4: Note that since the total population of patients and workloads served in both designed is constant, having both streaming designs under consideration as perfectly balanced entails $\rho'_{ij} = \rho/4$ and $\lambda'_{ij} = \lambda/4$ (for all $ij \in \mathbf{U} \times \mathbf{C}$). First consider the design where patients are streamed based on the complexity-based information (and prioritized based on the urgency-information). Label this design as D1. Using similar results to (5) and (6), and replacing λ with $\lambda/2$ (since the volume of patients in each stream are the same), for the stream of S patients under D1, ROAE denoted by R_S^{D1} is: $R_S^{D1} = \lambda \mathbb{E}(\bar{s}^2)/4 \left[\theta'_{US} \lambda'_{US} / (1 - \rho/4) + \theta'_{NS} \lambda'_{NS} / [(1 - \right.$

$\rho/4)(1 - \rho/2)]$. Similarly, for the stream of C patients under D1, ROAE denoted by R_C^{D1} is: $R_C^{D1} = \lambda \mathbb{E}(\tilde{s}^2)/4 \left[\theta'_{UC} \lambda'_{UC}/(1 - \rho/4) + \theta'_{NC} \lambda'_{NC}/[(1 - \rho/4)(1 - \rho/2)] \right]$. Furthermore, the total ROAE under D1 is $R^{D1} = R_S^{D1} + R_C^{D1}$. Next, consider the design where patients are streamed based on the urgency-based information (and prioritized based on complexity-based information). Label this design as D2, and notice that, for the stream of U patients under D2, ROAE denoted by R_U^{D2} is: $R_U^{D2} = \lambda \mathbb{E}(\tilde{s}^2)/4 \left[\theta'_{US} \lambda'_{US}/(1 - \rho/4) + \theta'_{UC} \lambda'_{UC}/[(1 - \rho/4)(1 - \rho/2)] \right]$. Similarly, for the stream of N patients under D2, ROAE denoted by R_N^{D2} is: $R_N^{D2} = \lambda \mathbb{E}(\tilde{s}^2)/4 \left[\theta'_{NS} \lambda'_{NS}/(1 - \rho/4) + \theta'_{NC} \lambda'_{NC}/[(1 - \rho/4)(1 - \rho/2)] \right]$. Also, the total ROAE under D2 is $R^{D2} = R_U^{D2} + R_N^{D2}$. Hence, after simplifications, we have:

$$R^{D1} - R^{D2} = \frac{\lambda \rho \mathbb{E}(\tilde{s}^2)}{8(1 - \rho/4)(1 - \rho/2)} [\theta'_{NS} \lambda'_{NS} - \theta'_{UC} \lambda'_{UC}],$$

which completes the proof after noticing that $\lambda'_{NS} = \lambda'_{UC}$ (as the streaming designs are assumed to be perfectly balanced), and (b) $\theta'_{NS} \leq \theta'_{UC}$ (as the optimal priority rules in each stream are assumed to be based on the Proposition 2(i) and the streaming designs are assumed to be perfectly balanced). \square

Online Appendix B (Comparison of Partially Balanced Streaming Designs)

In Section 5.3, we compared the performance of two streaming patient flow designs: streaming patients based on complexity information and prioritizing them based on urgency (complexity streaming), and streaming patients based on urgency and prioritizing them based on complexity (urgency streaming). We analytically showed that complexity streaming is preferred to urgency streaming when the streaming is done in a *perfectly balanced* manner. Here we consider the case where the streaming is only *partially balanced*; that is, we still assume that the utilizations in the two streams of each design can be balanced through appropriate capacity allocation but that volumes of patients sent to each stream, mean service times, service time variances will not be the same. We note that our collected data as well as some studies from the literature (e.g., Vance and Spirvulis (2005)) indicate that complexity can be defined in such a way that the volume of complex and simple patients are roughly equal. However, unlike complexity, in some ED's the percentage of urgent and non-urgent patients might not be roughly equal. Hence, we perform a sensitivity analysis on this factor. Because both the total population of patients and workloads served are constant in both designs, balancing the utilizations requires $\rho'_{ij} = \rho/4$ (for all $ij \in \mathcal{U} \times \mathcal{C}$). Next, while we keep the volume of patients sent to both streams in complexity streaming equal (i.e., $\lambda'_{US} + \lambda'_{NS} = \lambda'_{UC} + \lambda'_{NC}$), we assume that the volume of patients sent to the non-urgent stream is $(1 + \delta)$ times that of the urgent stream in urgency streaming and examine cases with $\delta \in \{-0.5, -0.25, 0, 0.25, 0.5\}$. After capacity allocation to make utilization equal in the two streams, the effective (error-impacted) service rates are $\mu'_{US} = \mu'_{UC} = 1/2$ and $\mu'_{NS} = \mu'_{NC} = (1 + \delta) \times 1/2$. We

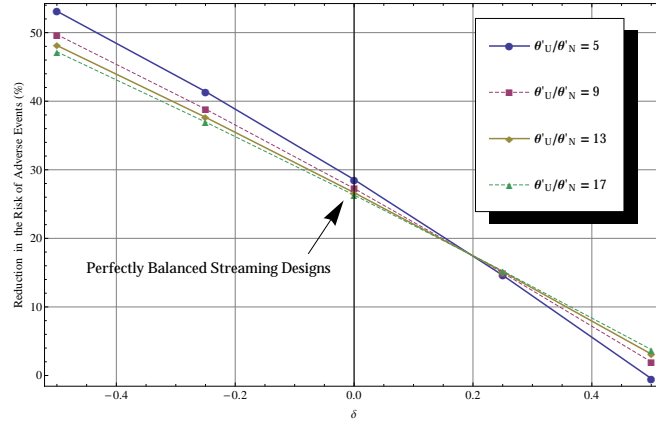


Figure 9 Comparison of complexity streaming with urgency streaming. $\delta \in [-0.5, 0.5]$ represents deviation from perfectly balanced streaming systems ($\theta'_{NC} = \theta'_{NS} = \theta'_N = 1, \theta'_{UC} = \theta'_{US} = \theta'_U$).

further assume that the effect of capacity allocation on the service time variances is roughly equal to square of mean service times. Figure 9 depicts the percentage benefit of complexity streaming over urgency streaming for a numerical example with a total ED utilization of 80% where (error-impacted) arrivals are $\lambda'_{US} = \lambda'_{UC} = 1/10$ and $\lambda'_{NS} = \lambda'_{NC} = (1 + \delta) \times 1/10$. As can be seen from Figure 9, complexity streaming is preferred to urgency streaming for these levels of δ over a wide range of values for the ratio of adverse event rates, θ'_U/θ'_N . Our other numerical examples and sensitivity analysis also show similar behavior, although we do not present them here for brevity.

Online Appendix C (General Occurrence of Adverse Events)

In the main body of the paper, we modeled the occurrence of adverse events with Poisson processes. We now relax the Poisson assumption and test the robustness of the Phase 1 priority rules proposed in the paper.

To this end, we allow the adverse events to occur based on any stationary point process (but not necessarily Poisson). For class $ij \in \mathcal{U} \times \mathcal{C}$, let S'_{ij} and $\{\Theta'_{ij}(t), t \geq 0\}$ denote the error-impacted random service time and the error-impacted process that defines the number of adverse events until time t , respectively. Also, recall that $R_{\pi}^{\mathcal{U} \cup \mathcal{C}'}(t)$ denotes the total number of adverse events until time t under policy π , and consider the following notions of stochastic ordering for two random variables X and Y . If $Pr\{X > u\} \geq Pr\{Y > u\}$ for all $u \in \mathbb{R}$, then X is said to be *stochastically* greater than Y ($Y \leq_{st} X$). Similarly, we say that the counting process $\{N_1(t), t \geq 0\}$ is *stochastically* greater than the counting process $\{N_2(t), t \geq 0\}$ ($\{N_2(t), t \geq 0\} \leq_{st} \{N_1(t), t \geq 0\}$) if for all $t \geq 0$ $N_2(t) \leq_{st} N_1(t)$. If f and g represent the densities or probability mass functions of X and Y , respectively, and $f(\xi)/g(\xi)$ is increasing in ξ over the union of the supports of X and Y , then X is said to be greater than Y in the *likelihood ratio ordering* ($Y \leq_{lr} X$) (see, e.g., Shaked and Shanthikumar (2007) for more details). Finally, we let Ξ denote the class of all static priority rules, and use a sample path argument to show the following.

PROPOSITION 5. *If classes $ij, i'j' \in \mathcal{U} \times \mathcal{C}$ are such that $\{\Theta'_{i'j'}(t), t \geq 0\} \leq_{st} \{\Theta'_{ij}(t), t \geq 0\}$ and*

$S'_{ij} \leq_{lr} S'_{i'j'}$, then a policy that prioritizes patients of class $i'j'$ over those of class ij (whenever patients of both classes are available) cannot be uniquely optimal in the sense of stochastically minimizing $R_{\pi}^{\mathcal{U}' \cup \mathcal{C}'}(t)$ at every $t \geq 0$ over all $\pi \in \Xi$.

Proof of Proposition 5: We use a sample path argument. To this end, we first state the following result, a more general version of which can be found in Righter 1994 (Lemma 13.D.1) among others.

LEMMA 3. Consider random variables X and Y , and let $\min\{X, Y\} = m$ and $\max\{X, Y\} = M$. If $X \leq_{lr} Y$, then $Pr\{X = m | m, M\} = Pr\{Y = M | m, M\} \geq Pr\{Y = m | m, M\} = Pr\{X = M | m, M\}$.

Suppose policy $\pi \in \Xi$ takes a patient from class $i'j'$ at time t_0 while a patient of class ij is waiting in the waiting area. We construct a new policy, $\hat{\pi} \in \Xi$, which takes a patient of class ij at time t_0 , and has a lower number of adverse events: $R_{\hat{\pi}}^{\mathcal{U}' \cup \mathcal{C}'}(t) \leq R_{\pi}^{\mathcal{U}' \cup \mathcal{C}'}(t)$ for all $t \geq 0$ along any sample path. To this end, using the above lemma, we employ a *cross coupling* (see, e.g., Section 13.D of Righter 1994) argument to prove the result as follows. Let $S'_l{}^{\gamma}$ be the “error-impacted” service time of patients of class $l \in \{ij, i'j'\}$ (i.e., the service time for those classified as l) under policy $\gamma \in \{\pi, \hat{\pi}\}$. Using Lemma 3 (and since $S'_{ij} \leq_{lr} S'_{i'j'}$), we can cross couple $(S'_{ij}{}^{\pi}, S'_{i'j'}{}^{\pi})$ with $(S'_{ij}{}^{\hat{\pi}}, S'_{i'j'}{}^{\hat{\pi}})$ so that $m \triangleq \min\{S'_{ij}{}^{\pi}, S'_{i'j'}{}^{\pi}\} = \min\{S'_{ij}{}^{\hat{\pi}}, S'_{i'j'}{}^{\hat{\pi}}\}$, $M \triangleq \max\{S'_{ij}{}^{\pi}, S'_{i'j'}{}^{\pi}\} = \max\{S'_{ij}{}^{\hat{\pi}}, S'_{i'j'}{}^{\hat{\pi}}\}$, and either $S'_{i'j'}{}^{\pi} = S'_{i'j'}{}^{\hat{\pi}} \triangleq \alpha \in \{m, M\}$ and $S'_{ij}{}^{\pi} = S'_{ij}{}^{\hat{\pi}} \triangleq \beta \in \{m, M\} \setminus \{\alpha\}$ (Case 1) or $S'_{ij}{}^{\pi} = S'_{ij}{}^{\hat{\pi}} = m$ and $S'_{i'j'}{}^{\pi} = S'_{i'j'}{}^{\hat{\pi}} = M$ (Case 2). For all other random variables, we use the usual “direct coupling” mechanism, i.e., we use the same realizations under both policies. For the occurrence of adverse events, since $\{\Theta'_{ij}(t), t \geq 0\}$ and $\{\Theta'_{i'j'}(t), t \geq 0\}$ are based on stationary point processes and $\{\Theta'_{ij}(t), t \geq 0\} \geq_{st} \{\Theta'_{i'j'}(t), t \geq 0\}$, we note that along any sample path we have $\Theta'_{ij}(t_1 + \xi) - \Theta'_{ij}(t_1) = \Theta'_{ij}(\xi) \geq \Theta'_{i'j'}(\xi) = \Theta'_{i'j'}(t_1 + \xi) - \Theta'_{i'j'}(t_1)$ for any fixed times $t_1, \xi \geq 0$ (note that stochastic dominance results in sample path dominance). Let $\tilde{t} > t_0$ be the time instance at which π takes a patient of class ij for the first time. Below, we consider cases 1 and 2 separately.

Case 1: In this case, the first decision epoch after t_0 is $t_0 + \alpha$. Policy $\hat{\pi}$ follows policy π for $t \in [t_0 + \alpha, \tilde{t})$. At \tilde{t} , π chooses a patient from class ij while $\hat{\pi}$ chooses a patient from class $i'j'$, and both policies finish serving them at $\tilde{t} + \beta$, at which the state of the system becomes the same under both policies, and $\hat{\pi}$ follows π from then on. Hence, we have $R_{\pi}^{\mathcal{U}' \cup \mathcal{C}'}(t) - R_{\hat{\pi}}^{\mathcal{U}' \cup \mathcal{C}'}(t) = (\Theta'_{ij}(t) - \Theta'_{ij}(t_0)) - (\Theta'_{i'j'}(t) - \Theta'_{i'j'}(t_0)) = \Theta'_{ij}(t - t_0) - \Theta'_{i'j'}(t - t_0) \geq 0$ for all $t \in [t_0, \tilde{t})$, and $R_{\pi}^{\mathcal{U}' \cup \mathcal{C}'}(t) - R_{\hat{\pi}}^{\mathcal{U}' \cup \mathcal{C}'}(t) = (\Theta'_{ij}(\tilde{t}) - \Theta'_{ij}(t_0)) - (\Theta'_{i'j'}(\tilde{t}) - \Theta'_{i'j'}(t_0)) = \Theta'_{ij}(\tilde{t} - t_0) - \Theta'_{i'j'}(\tilde{t} - t_0) \geq 0$ for all $t \geq \tilde{t}$ (and clearly $R_{\pi}^{\mathcal{U}' \cup \mathcal{C}'}(t) - R_{\hat{\pi}}^{\mathcal{U}' \cup \mathcal{C}'}(t) = 0$ for all $t < t_0$).

Case 2: In this case, $\hat{\pi}$ follows π at all decision epochs in $[t_0 + m, \tilde{t} + m - M)$ and serves a patient of class $i'j'$ at $\tilde{t} + m - M$ while π serves a patient of class ij at time \tilde{t} , the state under π and $\hat{\pi}$ becomes the same at $\tilde{t} + m$, and $\hat{\pi}$ follows π from then on. Thus, we have $R_{\pi}^{\mathcal{U}' \cup \mathcal{C}'}(t) - R_{\hat{\pi}}^{\mathcal{U}' \cup \mathcal{C}'}(t) = (\Theta'_{ij}(t) - \Theta'_{ij}(t_0)) - (\Theta'_{i'j'}(t) - \Theta'_{i'j'}(t_0)) = \Theta'_{ij}(t - t_0) - \Theta'_{i'j'}(t - t_0) \geq 0$ for all $t \in [t_0, t_0 + m)$, $R_{\pi}^{\mathcal{U}' \cup \mathcal{C}'}(t) - R_{\hat{\pi}}^{\mathcal{U}' \cup \mathcal{C}'}(t) \geq (\Theta'_{ij}(t) - \Theta'_{ij}(t_0)) - (\Theta'_{i'j'}(t) - \Theta'_{i'j'}(t_0)) = \Theta'_{ij}(t - t_0) - \Theta'_{i'j'}(t - t_0) \geq 0$

for all $t \in [t_0 + m, \tilde{t} + m - M)$ (where the first inequality holds since all the other patients in this interval are served under $\hat{\pi}$ no later than that under π), $R_{\pi}^{\mathcal{U}' \cup \mathcal{C}'}(t) - R_{\hat{\pi}}^{\mathcal{U}' \cup \mathcal{C}'}(t) \geq (\Theta'_{ij}(t) - \Theta'_{ij}(t_0)) - (\Theta'_{i'j'}(\tilde{t} + m - M) - \Theta'_{i'j'}(t_0)) \geq (\Theta'_{ij}(t) - \Theta'_{ij}(t_0)) - (\Theta'_{i'j'}(t) - \Theta'_{i'j'}(t_0)) = \Theta'_{ij}(t - t_0) - \Theta'_{i'j'}(t - t_0) \geq 0$ for all $t \in [\tilde{t} + m - M, \tilde{t})$, and $R_{\pi}^{\mathcal{U}' \cup \mathcal{C}'}(t) - R_{\hat{\pi}}^{\mathcal{U}' \cup \mathcal{C}'}(t) \geq (\Theta'_{ij}(\tilde{t}) - \Theta'_{ij}(t_0)) - (\Theta'_{i'j'}(\tilde{t} + m - M) - \Theta'_{i'j'}(t_0)) \geq (\Theta'_{ij}(\tilde{t}) - \Theta'_{ij}(t_0)) - (\Theta'_{i'j'}(\tilde{t}) - \Theta'_{i'j'}(t_0)) = \Theta'_{ij}(\tilde{t} - t_0) - \Theta'_{i'j'}(\tilde{t} - t_0) \geq 0$ for all $t \geq \tilde{t}$ (and clearly $R_{\pi}^{\mathcal{U}' \cup \mathcal{C}'}(t) - R_{\hat{\pi}}^{\mathcal{U}' \cup \mathcal{C}'}(t) = 0$ for all $t < t_0$). This completes the proof. \square

COROLLARY 1 (General Adverse Events). *If the four patient classes in $\mathcal{U} \times \mathcal{C}$ can be relabeled such that $\{\Theta'_{[4]}(t), t \geq 0\} \leq_{st} \{\Theta'_{[3]}(t), t \geq 0\} \leq_{st} \{\Theta'_{[2]}(t), t \geq 0\} \leq_{st} \{\Theta'_{[1]}(t), t \geq 0\}$ and $S'_{[1]} \leq_{lr} S'_{[2]} \leq_{lr} S'_{[3]} \leq_{lr} S'_{[4]}$, then prioritizing patients in the increasing order of their class labels stochastically minimizes $R_{\pi}^{\mathcal{U}' \cup \mathcal{C}'}(t)$ at every $t \geq 0$ for $\pi \in \Xi$.*

Proof of Corollary 1: The result follows directly from Proposition 5 using a simple interchange argument. \square

The above results presents a version of Proposition 2 (i) presented in the main body for the case where the Poisson assumption on the occurrence of adverse events is relaxed. The implication is that a similar priority rule remains optimal, provided that the complexity-based classification is defined in such a way that the “agreeable” ordering conditions of Corollary 1 hold.

Online Appendix D (The Clearing Model of Phase 2 ED Service)

Consider the simplified Phase 2 model with dynamic arrivals but, instead of assuming patients classified as $ij \in \mathcal{U} \times \mathcal{C}$ arrive according to a Poisson process with rate λ'_{ij} (superscript “’” indicates error-impacted rates), assume there are no arrivals but there are a finite number of patients already in the ED and available to the physician. The physician wants to clear the system in a manner that minimizes the expected number of adverse events. If the number of patients exceeds the number the provider can treat in parallel, then this scenario approximates backup overload periods.

To derive the optimal physician’s policy, we use the same notation used in the manuscript for the simplified Phase 2 model with dynamic arrivals. The following theorem shows that the priority rule presented in Theorem 1 remains optimal.

THEOREM 2 (Phase 2 Prioritization- Clearing). *In the simplified Phase 2 model with static arrivals, regardless of the number and class of available and unavailable patients, the physician should prioritize available patients in decreasing order of $p'_{ij} \hat{\theta}'_{ij} \hat{\mu}'_{ij}$. Furthermore, the physician should not idle when there is a patient available in an exam room.*

Proof of Theorem 2: Let $(\underline{x}, \underline{y})$ be the state of the system, where $\underline{x} = (x_{ij} : ij \in \mathcal{U} \times \mathcal{C})$ is the error-impacted number of the patients “available” to the physician and $\underline{y} = (y_{ij} : ij \in \mathcal{U} \times \mathcal{C})$ is the error-impacted number of the patients “unavailable” (i.e., under test, waiting for the results,

etc.). Also let $b < \infty$ denote the number of patients in the system at time zero (note that we have $(\underline{x}, \underline{y}) \mathbb{1}_{1 \times 8}^T \leq b$ at all time, where $\mathbb{1}$ denotes a vector of ones.). The goal is to move the system from a particular state $(\underline{x}, \underline{y})$ with $(\underline{x}, \underline{y}) \mathbb{1}_{1 \times 8}^T = b$ to the absorbing state $(\underline{0}, \underline{0})$, using a policy that results in the minimum expected number of adverse events. Note that since $(\underline{0}, \underline{0})$ is an absorbing state, the problem can be modeled as an infinite-horizon Markov Decision Process (MDP) with a discount factor $\beta \triangleq 1$ and a bounded transition rate $\psi = \max_{ij \in \mathbf{U} \times \mathbf{C}} \hat{\mu}'_{ij} + b\eta < \infty$.

After uniformization, the Bellman equation for any state in the state space $\mathcal{S} \triangleq \{(\underline{x}, \underline{y}) \in \mathbb{Z}_+^8 : (\underline{x}, \underline{y}) \mathbb{1}_{1 \times 8}^T \leq b\}$ can be written as follows (with the boundary condition $V(\underline{0}, \underline{0}) = 0$):

$$\begin{aligned} V(\underline{x}, \underline{y}) = & \frac{1}{\psi} \left[\hat{\theta}'(\underline{x} + \underline{y})^T + \beta \left[\sum_{ij \in \mathbf{U} \times \mathbf{C}} [y_{ij} \eta V(\underline{x} + \underline{e}_{ij}, \underline{y} - \underline{e}_{ij})] \right. \right. \\ & + \min_{a \in \mathcal{A}(\underline{x})} \left\{ \sum_{ij \in \mathbf{U} \times \mathbf{C}} \mathbb{1}_{\{a=ij\}} \hat{\mu}'_{ij} [p'_{ij} V(\underline{x} - \underline{e}_{ij}, \underline{y}) + (1 - p'_{ij}) V(\underline{x} - \underline{e}_{ij}, \underline{y} + \underline{e}_{ij})] \right. \\ & \left. \left. + (\psi - \sum_{ij \in \mathbf{U} \times \mathbf{C}} [y_{ij} \eta + \mathbb{1}_{\{a=ij\}} \hat{\mu}'_{ij}]) V(\underline{x}, \underline{y}) \right\} \right], \end{aligned} \quad (19)$$

where \underline{e}_{ij} is a vector with the same size as \underline{x} with a 1 in position ij and zeroes elsewhere, a is an action determining which patient class to serve, and $\mathcal{A}(\underline{x}) = \{ij \in \mathbf{U} \times \mathbf{C} : x_{ij} > 0\} \cup \{0\}$ is the set of feasible actions (class 0 represents the idling action which is considered as serving an extra class with all the rates equal to zero) when the error-impacted number of patients of each class who are "available" is \underline{x} .

To show the result, we use the finite-horizon equivalent optimality equation. When there are $n + 1$ (uniformized) periods to go, it can be written as:

$$\begin{aligned} V_{n+1}(\underline{x}, \underline{y}) = & \frac{1}{\psi} \left[\hat{\theta}'(\underline{x} + \underline{y})^T + \beta \left[\sum_{ij \in \mathbf{U} \times \mathbf{C}} [y_{ij} \eta V_n(\underline{x} + \underline{e}_{ij}, \underline{y} - \underline{e}_{ij})] \right. \right. \\ & + \min_{a \in \mathcal{A}(\underline{x})} \left\{ \sum_{ij \in \mathbf{U} \times \mathbf{C}} \mathbb{1}_{\{a=ij\}} \hat{\mu}'_{ij} [p'_{ij} V_n(\underline{x} - \underline{e}_{ij}, \underline{y}) + (1 - p'_{ij}) V_n(\underline{x} - \underline{e}_{ij}, \underline{y} + \underline{e}_{ij})] \right. \\ & \left. \left. + (\psi - \sum_{ij \in \mathbf{U} \times \mathbf{C}} [y_{ij} \eta + \mathbb{1}_{\{a=ij\}} \hat{\mu}'_{ij}]) V_n(\underline{x}, \underline{y}) \right\} \right]. \end{aligned} \quad (20)$$

We then use a convergence argument to prove the result for the original infinite horizon MDP. Grouping the terms related to control in the minimization and self-loop terms in (20), we can re-write it as:

$$\begin{aligned} V_{n+1}(\underline{x}, \underline{y}) = & \frac{1}{\psi} \left[\hat{\theta}'(\underline{x} + \underline{y})^T + \beta \left[\sum_{ij \in \mathbf{U} \times \mathbf{C}} [y_{ij} \eta V_n(\underline{x} + \underline{e}_{ij}, \underline{y} - \underline{e}_{ij})] \right. \right. \\ & - \max_{a \in \mathcal{A}(\underline{x})} \left\{ \sum_{ij \in \mathbf{U} \times \mathbf{C}} \mathbb{1}_{\{a=ij\}} \hat{\mu}'_{ij} [p'_{ij} \Delta_{ij}^y V_n(\underline{x} - \underline{e}_{ij}, \underline{y}) + \Delta_{ij}^{x,y} V_n(\underline{x} - \underline{e}_{ij}, \underline{y} + \underline{e}_{ij})] \right. \\ & \left. \left. + (\psi - \sum_{ij \in \mathbf{U} \times \mathbf{C}} y_{ij} \eta) V_n(\underline{x}, \underline{y}) \right\} \right], \end{aligned} \quad (21)$$

where $\Delta_{ij}^y V_n(\underline{x}, \underline{y}) = V_n(\underline{x}, \underline{y} + \underline{e}_{ij}) - V_n(\underline{x}, \underline{y})$ and $\Delta_{ij}^{x,y} V_n(\underline{x}, \underline{y}) = V_n(\underline{x} + \underline{e}_{ij}, \underline{y} - \underline{e}_{ij}) - V_n(\underline{x}, \underline{y})$.

To show the result, we use an *interchange* argument; we show that if classes $uc \in \mathbf{U} \times \mathbf{C}$ and $sl \in \mathbf{U} \times \mathbf{C}$ are such that $p'_{uc} \hat{\theta}'_{uc} \hat{\mu}'_{uc} \geq p'_{sl} \hat{\theta}'_{sl} \hat{\mu}'_{sl}$, then it is (weakly) better to serve class uc than

class sl when in state $(\underline{x}, \underline{y})$ with $x_{uc}, x_{sl} > 0$. This will also prove that the optimal policy will not idle the physician when there are one or more patients available in the rooms, since idling can be thought of serving an additional class, class 0, with $\hat{\theta}'_0 = \hat{\mu}'_0 = p'_0 = 0$

Let π ($\hat{\pi}$) be the policy that prescribes serving patients of class uc (sl) for every state $(\underline{x}, \underline{y}) \in \mathcal{S}$ with $x_{uc}, x_{sl} > 0$ and in every period n . Also, let $V_n^\pi(\cdot)$ and $V_n^{\hat{\pi}}(\cdot)$ denote the costs (when there are n periods to go) of policies π and $\hat{\pi}$, respectively. From (21), to show that π is (weakly) better than $\hat{\pi}$ in every period, we need to show that the following property holds for every n and every state $(\underline{x}, \underline{y}) \in \mathcal{S}$ with $x_{uc}, x_{sl} > 0$:

$$\begin{aligned} & \hat{\mu}'_{uc} [p'_{uc} \Delta_{uc}^y V_n^\pi(\underline{x} - \underline{e}_{uc}, \underline{y}) + \Delta_{uc}^{x,y} V_n^\pi(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc})] \\ & \geq \hat{\mu}'_{sl} [p'_{sl} \Delta_{sl}^y V_n^{\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y}) + \Delta_{sl}^{x,y} V_n^{\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y} + \underline{e}_{sl})]. \end{aligned} \quad (22)$$

To show property (22), we use induction on n . First, for $n = 0$, the property trivially holds since $V_0^\pi(\cdot, \cdot) = V_0^{\hat{\pi}}(\cdot, \cdot) = 0$. Next, suppose the property holds for n . We show that it will then also hold for $n + 1$. To do so, we need to consider different cases based on the state (i.e., partitions of the state space). First, if $b \geq 4$ consider the case where $x_{uc}, x_{sl} \geq 2$. Using action $a = uc$ (policy π) in both states $(\underline{x} - \underline{e}_{uc}, \underline{y})$ and $(\underline{x}, \underline{y})$ to compute $V_{n+1}^\pi(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc})$ and $V_{n+1}^\pi(\underline{x} - \underline{e}_{uc}, \underline{y})$ using (21), and subtracting the results we have $\Delta_{uc}^y V_{n+1}^\pi(\underline{x} - \underline{e}_{uc}, \underline{y}) =$

$$\begin{aligned} & \frac{1}{\psi_k} \left[\hat{\theta}'_{uc} + \beta \left[\sum_{ij \in \mathcal{U} \times \mathcal{C}} [y_{ij} \eta \Delta_{uc}^y V_n^\pi(\underline{x} + \underline{e}_{ij} - \underline{e}_{uc}, \underline{y} - \underline{e}_{ij})] \right. \right. \\ & \quad \left. \left. + \eta V_n^\pi(\underline{x}, \underline{y} - \underline{e}_{uc}) \right. \right. \\ & \quad \left. \left. - \hat{\mu}'_{uc} [p'_{uc} \Delta_{uc}^y \Delta_{uc}^y V_n^\pi(\underline{x} - 2\underline{e}_{uc}, \underline{y}) + \Delta_{uc}^y \Delta_{uc}^{x,y} V_n^\pi(\underline{x} - 2\underline{e}_{uc}, \underline{y} + \underline{e}_{ij})] \right. \right. \\ & \quad \left. \left. + (\psi - \sum_{ij \in \mathcal{U} \times \mathcal{C}} y_{ij} \eta) \Delta_{uc}^y V_n^\pi(\underline{x} - \underline{e}_{uc}, \underline{y}) - \eta V_n^\pi(\underline{x} - \underline{e}_{uc}, \underline{y}) \right] \right]. \end{aligned} \quad (23)$$

Similarly, we can derive $\Delta_{uc}^{x,y} V_{n+1}^\pi(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc})$ using (21) and action $a = uc$ (policy π) in both states $(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc})$ and $(\underline{x}, \underline{y})$ and subtracting the results. Doing so we have $\Delta_{uc}^{x,y} V_{n+1}^\pi(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc}) =$

$$\begin{aligned} & \frac{1}{\psi_k} \left[\beta \left[\sum_{ij \in \mathcal{U} \times \mathcal{C}} [y_{ij}^+ \eta \Delta_{uc}^{x,y} V_n^\pi(\underline{x} + \underline{e}_{ij} - \underline{e}_{uc}, \underline{y} - \underline{e}_{ij} + \underline{e}_{uc})] \right. \right. \\ & \quad \left. \left. - \eta V_n^\pi(\underline{x}, \underline{y}) \right. \right. \\ & \quad \left. \left. - \hat{\mu}'_{uc} [p'_{uc} \Delta_{uc}^{x,y} \Delta_{uc}^y V_n^\pi(\underline{x} - 2\underline{e}_{uc}, \underline{y} + \underline{e}_{uc}) + \Delta_{uc}^{x,y} \Delta_{uc}^{x,y} V_n^\pi(\underline{x} - 2\underline{e}_{uc}, \underline{y} + \underline{e}_{ij} + \underline{e}_{uc})] \right. \right. \\ & \quad \left. \left. + (\psi - \sum_{ij \in \mathcal{U} \times \mathcal{C}} [y_{ij}^+ \eta]) \Delta_{uc}^{x,y} V_n^\pi(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc}) + \eta V_n^\pi(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc}) \right] \right], \end{aligned} \quad (24)$$

where $y_{ij}^+ = y_{ij}$ for all $ij \neq uc \in \mathcal{U} \times \mathcal{C}$, and $y_{uc}^+ = y_{uc} + 1$. In a similar way, and by using action $a = sl$ (policy $\hat{\pi}$) in (21) quantities $\Delta_{sl}^y V_{n+1}^{\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y})$ and $\Delta_{sl}^{x,y} V_{n+1}^{\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y} + \underline{e}_{sl})$ can be computed. Next, to check property (22) for $n + 1$, multiply (23) by $p'_{uc} \hat{\mu}'_{uc}$, and (24) by $\hat{\mu}'_{uc}$ and add up the results. Similarly, multiply $\Delta_{sl}^y V_{n+1}^{\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y})$ and $\Delta_{sl}^{x,y} V_{n+1}^{\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y} + \underline{e}_{sl})$ by $p'_{sl} \hat{\mu}'_{sl}$ and $\hat{\mu}'_{sl}$, respectively,

and add up the results. Next, using the induction hypothesis and that $p'_{uc} \hat{\theta}'_{uc} \hat{\mu}'_{uc} \geq p'_{sl} \hat{\theta}'_{sl} \hat{\mu}'_{sl}$, after algebraic simplification it follows that

$$\begin{aligned} & \hat{\mu}'_{uc} [p'_{uc} \Delta_{uc}^y V_{n+1}^\pi(\underline{x} - \underline{e}_{uc}, \underline{y}) + \Delta_{uc}^{x,y} V_{n+1}^\pi(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc})] \\ & - \hat{\mu}'_{sl} [p'_{sl} \Delta_{sl}^y V_{n+1}^{\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y}) + \Delta_{sl}^{x,y} V_{n+1}^{\hat{\pi}}(\underline{x} - \underline{e}_{sl}, \underline{y} + \underline{e}_{sl})] \geq 0, \end{aligned} \quad (25)$$

which establishes property (22) for $n + 1$ for the case where $x_{uc}, x_{sl} \geq 2$. In a similar way, this property can be established for other cases (i.e., the remaining partition of the state space). Hence, the non-idling strict priority rule is optimal for all n . Next, taking the limit as $n \rightarrow \infty$ it follows that the finite horizon problem converges to the original infinite horizon one both in policy and cost (see, e.g., Sennott (1999) Proposition 4.3.1), which completes the proof. \square

References for the Online Appendices

- Buyukkoc, C., P. Varaiya, J. Walrand. 1985. The $c\mu$ rule revisited. *Adv. Appl. Prob.* 17, 237–238.
- Cox, D.R., W.L. Smith. 1961. *Queues*. Methuen & Co, London.
- Guo, X., K. Liu. 2001. A note on optimality conditions for continuous-time Markov Decision Processes with average cost criterion. *IEEE Trans. on Aut. Contr.* 46(12) 1984–1989.
- Kakalik, J.S., J.D.C. Little. 1971. *Optimal Service Policy for the M/G/1 Queue with Multiple Classes of Arrival*. Rand Corporation Report.
- Righter, R., *Scheduling, "Stochastic Orders and Their Applications"*, Chapter 13, edited by M. Shaked and J. G. Shanthikumar. New York: Academic Press, 381–432, 1994.
- Sennott, L.I. 1999. *Stochastic Dynamic Programming and the Control of Queueing Systems*. Wiley Series in Probability and Statistics, John Wiley and Sons, New York.
- Vance, J., P. Spirvulis. 2005. Triage nurses validly and reliably estimate Emergency Department patient complexity. *Emergency Medicine Australasia* 17, 382–386.