Working Paper

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

Wallace J. Hopp
Stephen M. Ross School of Business
at the University of Michigan

Soroush Saghafian
Dept. of Industrial & Operational Engineering
University of Michigan

Jeffrey S. Desmond
Emergency Department
University of Michigan Hospital

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
Nov 2011

Submitted to
manuscript XX-XXX

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

Soroush Saghafian[1], Wallace J. Hopp[2], Mark P. Van Oyen[1], Jeffrey S. Desmond[3] (M.D.), Steven L. Kronick[3] (M.D.)

[1] Dept. of Industrial & Operations Eng., Univ. of Michigan, Ann Arbor, MI
[2] Ross School of Business, Univ. of Michigan, Ann Arbor, MI
[3]Emergency Department, Univ. of Michigan Hospital, Ann Arbor, MI

Most hospital Emergency Departments (ED's) use triage systems that classify and prioritize patients almost exclusively on the basis of urgency. We demonstrate that the current practice of prioritizing patients solely based on urgency (e.g., ESI-2 patients over ESI-3 patients in the main ED) is less effective than a new ED triage system that adds an up-front estimate of patient complexity to the conventional urgency-based classification. Using a combination of analytic and simulation models calibrated with hospital data, we show that complexity-based triage can substantially improve both patient safety (i.e., reduce the risk of adverse events) and operational efficiency (i.e., shorten the average length of stay). Moreover, we find that ED's with high resource (physician and/or examination room) utilization, high heterogeneity between the average treatment time of simple and complex patients, and a relatively equal split between simple and complex patients benefit most from the proposed complexity-based triage system. Furthermore, while misclassification of a complex patient as simple is slightly more harmful than vice versa, complexity-based triage is robust to misclassification error rates as high as 25%. Finally, we show that up-front complexity information can be used to create two separate service streams, which facilitates the application of lean methods that amplify the benefit of complexity-based triage information.

Key words: Healthcare Operations Management; Emergency Department; Triage; Priority Queues; Patient prioritization; Markov Decision Processes.
History: Version: Nov., 2011.

1. Introduction

Overcrowding and lapses in patient safety are prevalent problems in Emergency Departments (ED's) in the U.S. and around the world. In one study, 91% of U.S. ED's responding to a national survey reported that overcrowding was a problem, and almost 40% of them reported overcrowding as a daily occurrence (American Hospital Association (2002)). In addition to causing long wait times, many research studies have linked delays due to overcrowding to elevated risks of errors and adverse events (see, e.g., Thomas et al. (2000), Gordon et al. (2001), Trzeciak and Rivers (2003), and Liu et al. (2005)). This situation prompted the Institute of Medicine's Committee on Future of Emergency Care in the United States Health System to recommend that *"hospital chief executive officers adopt enterprisewide operations management and related strategies to improve the quality and efficiency of emergency care"* (Institute of Medicine (2007)). The triage process is a natural

place to introduce operations management (OM) into the ED.

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) argue 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 Iserson (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 other U.S. hospitals have adopted the 5-level Emergency Severity Index (ESI) system (see Fernandes et al. (2005)), which combines urgency with an estimate of resources (e.g., tests) required. In the ESI system (a typical version of which is illustrated in Figure 1 (left)), urgent patients who cannot wait are classified as ESI-1 and 2, while non-urgent patients who can wait are classified as ESI-3, 4, and 5. ESI-4 and 5 patients are usually directed to a fast track (FT) area, while ESI-1 patients are immediately moved to a resuscitation unit (RU). ESI-2 and 3 patients, who represent the majority of patients at large academic hospitals (e.g., about 80% at the University of Michigan ED (UMED)), are served in the main area of the ED with priority given to ESI-2 patients. Since the ESI system does not differentiate between patients in the ESI-2 and ESI-3 categories in terms of complexity, patients in the main ED are still sorted and prioritized purely on the basis of urgency. Hence, the ESI system does not respond to the second purpose of triage for the majority of the patients. As Welch and Davidson (2011) state, *"Many clinicians have already realized that triage as it is widely*

practiced today no longer meets the requirement of timely patient care." Our goal in this paper is to propose a new triage system, which we call *complexity-based triage*, that can significantly improve ED performance with respect to both clinical justice and efficiency.

Doing this poses two challenges: (a) deciding what information should be collected at the time of triage, and (b) determining how this information should be used to assign patients to tracks and prioritize them within tracks (see, e.g., King et al. (2006)). Saghafian et al. (2010) proposed that one way ED's can improve performance is to have triage nurses predict the final disposition (admit or discharge) of patients in addition to assigning an ESI level. 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 this study 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. This suggests that classifying patients according to complexity may be even more useful than classifying them according to ultimate disposition.

There is ample evidence from the OM literature that classifying patients based on their service requirements and giving priority to those with shorter service times (e.g., by following a Shortest Processing Time (SPT) priority rule) can improve resource usage efficiency, and thereby reduce the average waiting time among all patients. Furthermore, empirical studies from the emergency medicine literature suggest that patients can be effectively classified by complexity at the time of triage. Specifically, Vance and Spirvulis (2005) defined complex patients as those requiring at least two procedures, investigations, or consultations and concluded that *"Triage nurses are able to make valid and reliable estimates of patient complexity. This information might be used to guide ED work flow and ED casemix system analysis."*

Using the number of (treatment related) interactions with the physician (which correlates directly with expected treatment duration) as an indicator of patient complexity, we propose and investigate the benefit of the new complexity-based triage process depicted in Figure 1 (right). Note that, unlike the ESI system, our proposed system classifies all patients (except those at risk of death) in

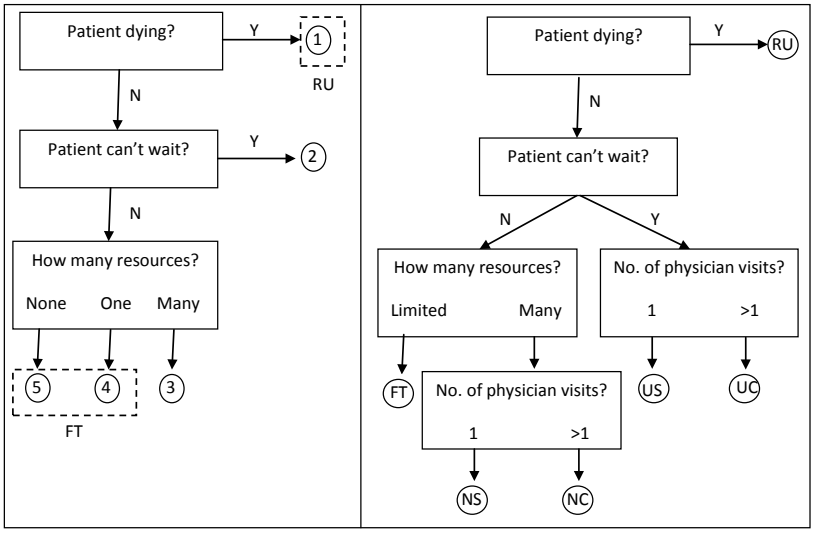


Figure 1 Left: Current practice of triage (Emergency Severity Index (ESI) algorithm version 4); Right: Proposed complexity-based triage system (RU: Resuscitation Unit, FT: Fast Track, NS: Non-urgent Simple, NC: Non-urgent Complex, US: Urgent Simple, UC: Urgent Complex).

terms of complexity. In this paper, we compare our proposed triage system with current urgency-based systems and show that incorporating patient complexity into the triage process can yield substantial performance benefits. To do this, we consider ED performance in terms of both risk of adverse events (clinical justice) and average length of stay (efficiency). Specifically, we make use of a combination of analytic and simulation models calibrated with hospital data to examine the following:

1. Prioritization: *How should ED's use complexity-based triage information to prioritize patients?*

2. Magnitude: *How much benefit does complexity-based triage (which adds complexity information to conventional urgency evaluations) offer relative to urgency-based triage?*

3. Sensitivity: *How sensitive are the benefits of complexity-based triage to misclassification errors and other characteristics that may vary across ED's?*

4. Design: *Should complexity-based information be used to create separate service streams for simple and complex patients, or is it better to use it to prioritize patients in a traditional pooled flow design?*

In addition to collecting detailed ED data (from UMED), addressing these practical questions required us to make some technical innovations: (1) In the ED, upfront triage misclassifications are inevitable. However, the literature on priority queueing systems under misclassification is very lim-

ited. We contribute to this literature by explicitly considering misclassifications and deriving optimal control policies under different settings that effectively approximate the ED environment. We do this through a linear transformation of control indices so that they represent "error-impacted" rates, which use only information from historical data. This leads to modified versions of the well-known $c\mu$ rule, which we show to be very effective as the basis for prioritizing patients into ED examination rooms. (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 heath delivery settings, is that patients are occasionally 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 (controller) may need to consider both the current and the future availability of the patients when making decisions. This type of problems usually result 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 MDP model of patient prioritization within examination rooms cannot use the conventional method of uniformization (proposed by Lippman (1975)) for working with continuous-time MDP's. The available technical results for continuous-time MDP's with unbounded transition rates is very limited (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. Using this innovative technique, we derive a simple-to-implement rule for ED physicians that prescribes which patient to visit next.

The remainder of the paper is organized as follows. Section 2 summarizes previous OM and medical 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 focuses on Phase 1 and uses analytical queueing models to compare performance under urgency-based

and complexity-based triage systems. Section 5 considers Phase 2 by developing and analyzing a Markov Decision Process model. Section 6 uses a high-fidelity simulation model of the full ED to validate the insights obtained through our analytical models and to refine our estimates of the magnitude of performance improvement possible with complexity-based triage. We conclude in Section 7.

2. Literature Review

In this section, we review studies related to our work from both the operations research/management literature and the medical literature.

2.1. Operations Research/Management Studies

The effect of assigning priorities in queueing systems has been studied in the operations research literature for a long time. One of the first works to rigorously analyze such systems under perfect classification was Cobham (1954). Assuming perfect customer classification, Cobham (1954) and Cobham (1955) showed that the expected waiting time among all customers can be reduced by assigning priorities. van der Zee and Theil (1961) extended Cobham's results to the case with imperfect classification for a two-priority single-channel system. They recommended creating a "mixed" group for customers who cannot be classified with certainty into either group 1 or 2, and assigning priorities probabilistically within this group. Further analysis of priority queueing systems can be found in Cox and Smith (1961), Jaiswal (1968), and Wolf (1989).

Under perfect classification, average holding cost objective, Poisson arrivals, and a non-preemptive non-idling single server model, Cox and Smith (1961) used an interchange argument to show that the $c\mu$ rule is optimal among priority rules. That is, product of the holding cost rate times the service rate is the index that quantifies the attractiveness/priority of that job or job class. Kakalik and Little (1971) extended this result and used a semi-Markov decision process 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), Veatch (2010),

Saghafian et al. (2011), and references therein. In this paper, we contribute to this literature by proving the optimality of modified versions of the $c\mu$ rule that use "error-impacted" indices, which are well-suited to the ED triage environment where misclassification is inevitable.

Research related to our work that analyzes the performance of ED's from an operations perspective is also very limited. Saghafian et al. (2010) considered streaming of ED patients based on triage estimations of the final disposition (admit or discharge) and found that an appropriate "virtual streaming" policy can improve performance with respect to the operational characteristics of average Length of Stay (LOS) and Time To First Treatment (TTFT). 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 heterogenous high risk patients, for which treatment times are exponential, and patient classification is perfect, and concluded that patients should be prioritized into as many urgency classes as possible in order to maximize survival. Argon and Ziya (2009) used the 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.

The above studies suggest that separating patients according to a measure of service duration can reduce waiting times through a better resource allocation. However, we note that they (a) lack insights into clinical justice/safety issues that are vital in ED's, and (b) are limited to simple/stylized queueing models with features (e.g., one-stage service, fixed number of customers all available at time zero, availability of the customers at any time during service, no bound on the number of customers that can be assigned to a server, no change in the condition/holding cost of customers after they begin service, perfect customer classification, etc.) that do not capture the reality of ED's. In this paper, we seek to address both safety and efficiency, and to account for the key features that define the ED environment. To this end, in addition to using stylized models that

approximate ED flow, we develop a complex simulation model of the ED and use hospital data to investigate whether the insights from stylized models carry over to the actual ED environment.

2.2. Medical Studies

Our research was informed by empirical studies of ED's and triage processes. 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. Most studies attribute the first formal battlefield triage system to the distinguished French military surgeon Baron Dominique-Jean Larrey who recognized a need to evaluate and categorize wounded soldiers. He recommended treating and evacuating those requiring the most urgent medical attention, rather than waiting hours or days for the battle to end before treating patients, as had been done in previous wars (Ierson and Moskop (2007)). Since that time, triage in medicine has been mainly based on urgency. However, the idea of considering the complexity of patients goes back to World War I 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)). The ESI triage system shown in Figure 1 (left) is the most serious effort to date at introducing complexity into the triage process. However, because (a) the number of resources required does not necessarily correlate with the physician time required by the patient, (b) The complexity of patients varies greatly within ESI categories, and (c) ED's do not use ESI information in a consistent manner to prioritize patients, the ESI system falls well short of the potential for complexity-based triage.

Anticipating the potential of complexity-based 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 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. Other researchers have suggested using physicians at triage as a way to generate more and better patient information. However, Han et al. (2010) and Russ et al. (2010) studied physician triage and found that it is

not an effective method for reducing total length of stay, although it may reduce the average time spent in an ED bed.

Finally, several studies have been published in medical journals that aim at investigating and/or validating the ESI triage system. For this stream of research, we refer interested readers to Fernandes et al. (2005), which summarized the findings and recommendations of a task force from the American College of Emergency Physicians (ACEP) and the Emergency Nurses Association (ENA) appointed in 2003 to analyze the literature and make recommendations regarding use of 5-level triage systems in the United States. While this committee found the 5-level ESI system to be a good option compared to other available methods, they encouraged further in-depth research for improving the triage system.

3. Modeling the ED

To answer the four questions (prioritization, magnitude, sensitivity, and design) we posed in Section 1, we need to model patient flow through the ED. A high level schematic of this flow is presented in Figure 2. A patient's path through the 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 ED *waiting* area until s/he is called by the charge nurse and brought to an examination room. There s/he goes through a stochastic number of *treatment* stages with a physician, which are also stochastic in duration. These treatment stages are punctuated by *test* stages which involve testing (MRI, CT Scan, X-Ray etc.) or preparation/processing activities that do not involve the physician and during which the patient is unavailable to the physician. 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 remainder of time in the ED until disposition as "Phase 2."

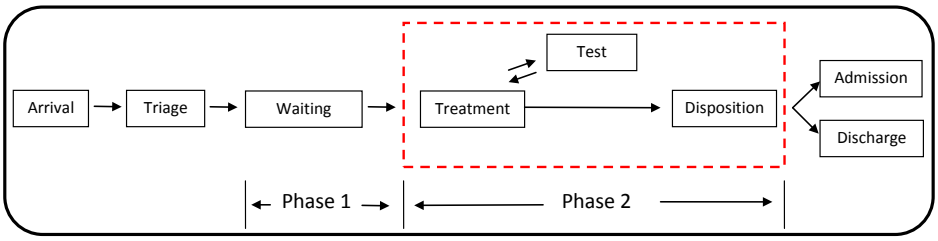


Figure 2 General flow of patients in the main ED.

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 can make use of the patient classification performed at triage. Similarly, in Phase 2, physicians use some kind of a sequencing rule to choose which patient to see next.

To gain insights into appropriate triage and priority rules, we first focus on the risk of adverse events and average waiting times in Phase 1 by considering the dashed area in Figure 2 (i.e., Phase 2) as a single-stage service node with a single, aggregated server. Since ED's rarely send a patient back to the waiting area of Phase 1 once s/he has begun service, we assume a non-preemptive service protocol. We also approximate the non-stationary arrival process by a stationary Poisson process. These simplifications allow us to gain insights into suitable Phase 1 priority rules using a multi-class non-preemptive priority M/G/1 queueing model. We refer to this model as the *simplified single-stage ED model*. An important and challenging aspect of this model is the existence of triage misclassifications that can affect the way patients should be prioritized.

After analyzing this model, we focus on the risk of adverse events and average waiting times in Phase 2. To do this, we note that physicians can preempt their current interaction with a patient to visit another patient with a higher priority (e.g., a severely acute patient), and hence, we allow for preemption in Phase 2. Again approximating arrivals with a stationary Poisson process arrival stream, we can represent the multi-stage service process in Phase 2 as a Markov decision process model, which we label as the *simplified multi-stage ED model*. We use this model to get insights into appropriate Phase 2 priority rules that physicians can implement when choosing their next patient.

Finally, we test the insights from both analytic models under realistic conditions with a high fidelity simulation model of the full ED calibrated with a year of data from University of Michigan Hospital ED as well as time study data from the literature.

4. Phase 1: A Simplified Single-Stage ED Model

To formalize the Phase 1 sequencing problem, we define a patient to be of type ij if his/her urgency level is $i \in \boldsymbol{\mathcal{U}}$ and his/her complexity type is $j \in \boldsymbol{\mathcal{C}}$, where $\boldsymbol{\mathcal{U}} = \{U(\text{Urgent}), N(\text{Non-urgent})\}$ and $\boldsymbol{\mathcal{C}} = \{C(\text{Complex}), \text{S}(\text{Simple})\}$. We suppose patients of type $ij \in \boldsymbol{\mathcal{U}} \times \boldsymbol{\mathcal{C}}$ arrive according to a Poison process with rate λ_{ij} and have service times (i.e., the total time spent in Phase 2) that follow a distribution, $F_{ij}(s)$ with first moment $1/\mu_{ij}$ (where $\mu_{iC} \leq \mu_{iS}$ for all $i \in \boldsymbol{\mathcal{U}}$) and a finite second moment. We assume patients of type ij are subject to adverse events which occur according to a Poisson process with rate θ_{ij}, where $\theta_{Uj} \geq \theta_{Nj}$ for all $j \in \boldsymbol{\mathcal{C}}$. Notice that adverse events only rarely result in death, i.e., the average reported number of adverse events per patient is much higher than the average number of death per patient (see, e.g., Liu et al. (2005) where the authors report that 28% of patients boarded in the ED had some adverse event or error in the course of boarding only). Thus, we assume that the service process continues, so that it is possible for a patient to experience more than one adverse event. This allows us to compare the performance of the ED under different triage systems in a systematic way. Similarly, changes in patient priority after the occurrence of an adverse event can be neglected, since (a) such changes are rare, and (b) the effect of such rare changes are not systematically different under different triage systems.

Assuming $R_\pi^\Omega(t)$ represents the counting process that, under patient classification (i.e., triage) policy Ω and sequencing rule π, counts the total number of adverse events (for all patients) until time t, we consider $R_\pi^\Omega = \lim_{t\to\infty} R_\pi^\Omega(t)/t$ (when the limit exists) as our metric and refer to it as the *rate of adverse events (ROAE)*. However, if $\theta_{ij} = 1$ for all $i \in \boldsymbol{\mathcal{U}}$ and $j \in \boldsymbol{\mathcal{C}}$, then it can be shown that $R_\pi^\Omega / \sum_{i\in\boldsymbol{\mathcal{U}}} \sum_{j\in\boldsymbol{\mathcal{C}}} \lambda_{ij}$ represents the average *length of stay (LOS)*. (Notice that the sample path costs of LOS and adverse events with unit risk rates divided by total arrival rate will be different, but they are equal in expectation.) Hence, we can use our metric to characterize performance with respect to both safety and efficiency in a systematic and coherent way.

4.1. Urgency-Based Triage - Phase 1

We first consider current practice in most ED's, in which patients are classified solely based on urgency, and use our simplified single-stage model to focus on Phase 1 sequencing decisions. 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 $\lambda_U = \sum_{j\in\boldsymbol{\mathcal{C}}} \lambda_{Uj}$ and $\lambda_N = \sum_{j\in\boldsymbol{\mathcal{C}}} \lambda_{Nj}$, respectively. Similarly, the average service times for U's and N's are $1/\mu_U = \sum_{j\in\boldsymbol{\mathcal{C}}}(\lambda_{Uj}/\lambda_U)(1/\mu_{Uj})$ and $1/\mu_N = \sum_{j\in\boldsymbol{\mathcal{C}}}(\lambda_{Nj}/\lambda_N)(1/\mu_{Nj})$, respectively. Furthermore, from known results for non-preemptive priority queues (see, Cobham (1954), van der Zee and Theil (1961), Section 3.3 of Cox and Smith (1961), or Section 10.2 of Wolf (1989)), the average waiting (queue) time of the kth priority class is

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

where $\rho_l = \lambda_l/\mu_l$ 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 rate 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., patient classification with respect to set $\boldsymbol{\mathcal{U}}$) that gives priority to U's is

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

Similarly, we can obtain the ROAE under an urgency-based triage policy that gives priority to N's:

$$R_N^{\boldsymbol{\mathcal{U}}} = \theta_N\lambda_N(\lambda\mathbb{E}(s^2)/2(1-\rho_N)) + \theta_U\lambda_U(\lambda\mathbb{E}(s^2)/2(1-\rho_N)(1-\rho)). \tag{3}$$

Comparing these reveals that, without misclassification errors, the best priority rule is to prioritize U's (N's) if, and only if, $\theta_U\mu_U \ge (\le)\theta_N\mu_N$. 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 carefully considers the effect of stochastic triage misclassifications.

Therefore, we now formally incorporate stochastic misclassification errors into our decision model of urgency-based triage and prioritization. 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 ROAE 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 (2) with these new "error impacted" rates shows that when priority is given to U's, the ROAE under imperfect classification is:

$$R_U^{\boldsymbol{\mathcal{U}}'} = \theta'_U\lambda'_U(\lambda\mathbb{E}(s^2)/2(1-\rho'_U)) + \theta'_N\lambda'_N(\lambda\mathbb{E}(s^2)/2(1-\rho'_U)(1-\rho)), \tag{4}$$

where $\rho'_U = \lambda'_U/\mu'_U$. Similarly, using (3) shows that when priority is given to N's:

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

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

The above results enable us to state:

PROPOSITION 1 **(Phase 1 Prioritization - Urgency-Based Triage)**. *In the simplified single-stage ED model with imperfect urgency-based classification:*

(i) The best priority rule is to prioritize U patients if $\theta'_U\,\mu'_U \geq \theta'_N\mu'_N$; otherwise, prioritize N patients.

(ii) The best 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.

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, the above proposition

implies that doing so remains optimal even under realistic levels of misclassification errors. Hence, prioritizing ESI-2 patients over ESI-3 patients in the main ED seems legitimate in the current urgency-based triage practice in the U.S. However, we note that there is wide variance of complexity among ESI-2 and ESI-3 patients. Hence, if complexity is taken into account, simply prioritizing ESI-2 patients over ESI-3 patients as is currently done in practice for these majority of ED patients may be significantly suboptimal. We investigate this issue in the next section.

4.2. Complexity-Based Triage - Phase 1

We now consider the complexity-based triage policy shown in Figure 1 (right), and compare its performance with respect to that of urgency-based triage currently in use in practice. By doing this we seek to gain insights into the prioritization, magnitude, and sensitivity questions posed in the Introduction.

To evaluate the performance of complexity-based triage when classification is imperfect, we let γ_U and γ_N denote the misclassification error rates with respect to set $\boldsymbol{\mathcal{U}}$. That is, γ_U and γ_N denote the probabilities of classifying a U patient as an N, and an N patient as a U, respectively. Similarly, let γ_C and γ_S denote the misclassification error rates with respect to set $\boldsymbol{\mathcal{C}}$; γ_C denotes the probability that a C patient is classified as an S, and γ_S denotes the probability that an S patient is classified as a C. We assume the misclassification probabilities with respect to sets $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{C}}$ are independent. 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) have tested the ability of triage nurses to evaluate patient complexity (where complexity is defined as requiring two or more procedures, investigations, or consultation) and observed a misclassification rate of 17%.

Similar to what we did in Section 4.1, 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

$$A = \begin{pmatrix} (1-\gamma_U)(1-\gamma_S) & (1-\gamma_U)\gamma_C & \gamma_N(1-\gamma_S) & \gamma_N\gamma_C \\ (1-\gamma_U)\gamma_S & (1-\gamma_U)(1-\gamma_C) & \gamma_N\gamma_S & \gamma_N(1-\gamma_C) \\ \gamma_U(1-\gamma_S) & \gamma_U\gamma_C & (1-\gamma_N)(1-\gamma_S) & (1-\gamma_N)\gamma_C \\ \gamma_U\gamma_S & \gamma_U(1-\gamma_C) & (1-\gamma_N)\gamma_S & (1-\gamma_N)(1-\gamma_C) \end{pmatrix}. \quad (6)$$

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 multiplier and division, respectively.

With these, the waiting times for each customer class under an imperfect $\boldsymbol{\mathcal{U}} \cup \boldsymbol{\mathcal{C}}$ classification can be computed using (1) with rates replaced with their transformed error impacted counterparts. This model permits us to show the following.

PROPOSITION 2 **(Phase 1 Prioritization - Complexity-Based Triage)**. *In the simplified single-stage ED 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) $R_*^{\boldsymbol{\mathcal{U}}' \cup \boldsymbol{\mathcal{C}}'} \leq R_*^{\boldsymbol{\mathcal{U}}'}$. That is, even with misclassification errors, implementing the best priority rule for complexity-based triage is always (weakly) better than the optimal priority rule for urgency-based triage.

(iii) The best priority 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. Proposition 2 (ii) begins to address the magnitude question by suggesting that complexity-based triage outperforms urgency-based triage, given that the optimal priority rule is implemented. While priority rules are greedy and usually suboptimal, part (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. Furthermore, 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).

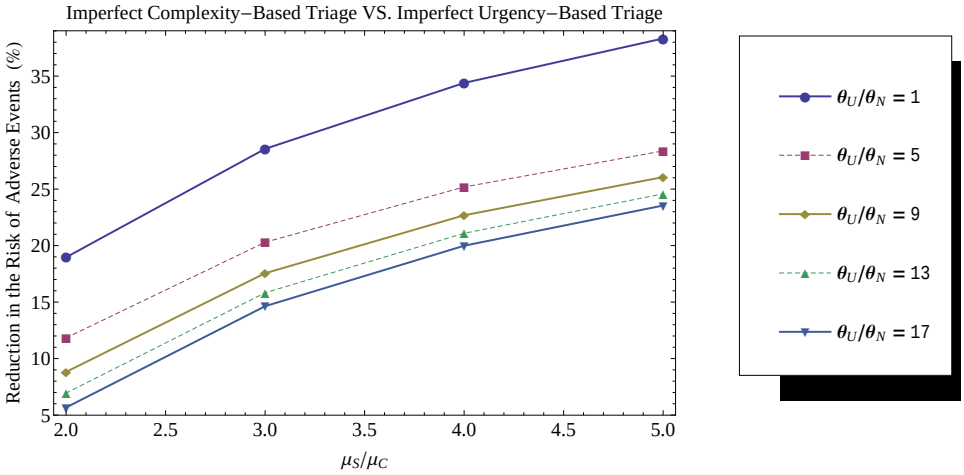


Figure 3 Benefit of complexity-based triage over urgency-based triage with practical misclassification rates reported in the literature ($\gamma_{\mathbf{U}} = \gamma_{\mathbf{N}} = \mathbf{10}\%, \gamma_{\mathbf{S}} = \gamma_{\mathbf{C}} = \mathbf{17}\%$).

Figure 3 provides additional insights into the magnitude question by illustrating the amount of improvement for a numerical example with $\mu_{UC} = \mu_{NC} = \mu_C = 1$, $\mu_{US} = \mu_{NS} = \mu_S$ varying from 2 to 5, $\lambda_{US} = (1/5)\mu_{US}, \lambda_{UC} = 1/4, \lambda_{NS} = (1/3)\mu_{NS}, \lambda_{NC} = 1/6, \mathbb{E}(s^2) = 4, \theta_{NS} = \theta_{NC} = \theta_N = 1, \theta_{US} = \theta_{UC} = \theta_U$. Note that (1) the amount of improvement is depicted both in terms of average length of stay and risk of adverse events (since when $\theta_U/\theta_N = 1$, the percentage improvement in risk of adverse events and length of stay are equal), and (2) reduction in the length of stay results in reduction in congestion (by Little's Law), which can serve as a potent remedy for the prevalently observed phenomenon of ED overcrowding. Figure 3 suggests that, if the average service time of complex patients is 3–4 times larger than that of simple patients, then complexity-based triage can reduce the risk of adverse events (ROAE) and average length of stay (LOS) by 12–22% and 27–33%, respectively. Finally, we can address the sensitivity question by using our model to determine the environmental factors that favor complexity-based triage.

PROPOSITION 3 **(Attractiveness of Complexity-Based Triage)**. *Under the simplified single-stage ED model, complexity-based 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) a more equal fraction of simple and complex patients, and (iv) lower error rates in classifying simple and complex patients.*

5. Phase 2: A Multi-Stage ED Model

The analysis of the previous section was limited to patient waiting and risk of adverse events prior to entry into an examination room (i.e., Phase 1). But, as illustrated in Figure 4, a great deal of ED activity takes place after this point, which contributes to both patient length of stay and risk of adverse events. Since triage classification can be used to sequence patients within the ED, as well as in the waiting room, it is important to consider Phase 2 sequencing as part of our evaluation of complexity-based triage. It seems intuitive that a priority rule similar to that for Phase 1 should serve as a useful guide to physicians in allocating their time among their slate of patients. However, investigating this requires a model of Phase 2 that includes some challenging new features (e.g., patients going off for tests, multiple patients of different class in different stages of treatment, etc.), which did not exist in the model of Phase 1.

To formulate a suitable model, consider the multi-stage service process illustrated in Figure 4 and suppose patients of class $ij \in \boldsymbol{\mathcal{U}} \times \boldsymbol{\mathcal{C}}$ arrive according to a Poison process with rate λ_{ij}. Further, we suppose the rate of adverse events in Phase 2 is denoted by the vector $\underline{\hat{\theta}} = (\hat{\theta}_{ij})_{ij \in \boldsymbol{\mathcal{U}} \times \boldsymbol{\mathcal{C}}}$ (which is usually less than the risk of adverse events in Phase 1, $\underline{\theta}$, because patients are monitored and treated 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. We call patients who have completed tests (e.g., MRI, CT Scan, X-Ray, etc.) and have results and are ready for a physician visit "available," and patients being tested, prepared, or waiting for results "unavailable." Our model for the congestion for tests is exogenous and can be estimated from historical data. An important feature in modeling the physician choice is the uncertain duration of unavailability of patients to the physician due to the wait for tests and their results.

Suppose each interaction with a patient of class ij takes an exponentially distributed amount of time with rate $\hat{\mu}_{ij}$ and assume (for tractability) 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 a realistic simulation model.)

After each completed interaction, a patient of class 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, it 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 class $ij \in \boldsymbol{\mathcal{U}} \times \boldsymbol{\mathcal{C}}$. Furthermore, we model test times, which include any preparation and wait times associated with the test, as a $\cdot/M/\infty$ queueing system with average service time of η^{-1}. Because we aggregate test times, waiting times for the test results, and preparations for tests into a single "test" stage, and also aggregate these for all possible types of tests, the long-run average time spent for a generic "test," denoted by η^{-1}, can be assumed to be roughly similar among different patient classes (for more detailed data on test turnaround times see Steindel and Howanitz (1997) and Holland et al. (2005)).

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 $\underline{x} = (x_{ij})_{ij \in \boldsymbol{\mathcal{U}} \times \boldsymbol{\mathcal{C}}}$ (respectively $\underline{y} = (y_{ij})_{ij \in \boldsymbol{\mathcal{U}} \times \boldsymbol{\mathcal{C}}}$) represent the 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 of 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 stablizable; i.e., there exists at least one policy under which the risk of adverse events is finite (otherwise, the problem does

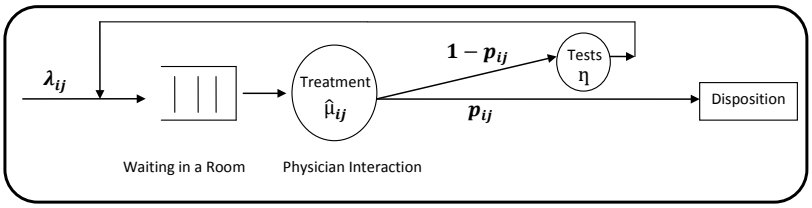


Figure 4 Patient flow after a patient is moved to an examination room/bed (Phase 2 sequencing).

not represent a real ED). However, notice that since the transition rates are not bounded, we cannot use uniformization in the spirit of Lippman (1975) to formulate a discrete time equivalent of the CTMC where the times between consecutive events are i.i.d. (for all states). However, in what follows, 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 $k\to\infty$. 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 as opposed to the state space, thereby avoiding the artificial boundary effects that usually impact the policy. Since the transition rates in the CTMC indexed by k (for all k) are bounded by $\psi_k=\max_{ij\in\boldsymbol{\mathcal{U}}\times\boldsymbol{\mathcal{C}}}\hat{\mu}_{ij}+4\,k\,\eta+\sum_{ij\in\boldsymbol{\mathcal{U}}\times\boldsymbol{\mathcal{C}}}\lambda_{ij}<\infty$, 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 $k\to\infty$) 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, R^{k*}, and the optimal physician behavior can be derived from the following average cost optimality equation:

$$\begin{aligned}J^k(\underline{x},\underline{y})+R^{k*}=\frac{1}{\psi_k}\Big[&\hat{\underline{\theta}}\,(\underline{x}+\underline{y})^T+\sum_{ij\in\boldsymbol{\mathcal{U}}\times\boldsymbol{\mathcal{C}}}\left[\lambda_{ij}J^k(\underline{x}+\underline{e}_{ij},\underline{y})+(y_{ij}\wedge k)\,\eta\,J^k(\underline{x}+\underline{e}_{ij},\underline{y}-\underline{e}_{ij})\right]\\&+\min_{a\in\mathcal{A}(\underline{x})}\Big\{\sum_{ij\in\boldsymbol{\mathcal{U}}\times\boldsymbol{\mathcal{C}}}\mathbb{1}_{\{a=ij\}}\hat{\mu}_{ij}\left[p_{ij}\,J^k(\underline{x}-\underline{e}_{ij},\underline{y})+(1-p_{ij})J^k(\underline{x}-\underline{e}_{ij},\underline{y}+\underline{e}_{ij})\right]\\&+\big(\psi_k-\sum_{ij\in\boldsymbol{\mathcal{U}}\times\boldsymbol{\mathcal{C}}}\left[\lambda_{ij}+(y_{ij}\wedge k)\,\eta+\mathbb{1}_{\{a=ij\}}\hat{\mu}_{ij}\right]\big)\,J^k(\underline{x},\underline{y})\Big\}\Big],\end{aligned}\tag{7}$$

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 $\mathcal{A}(\underline{x}) = \{ij \in \boldsymbol{\mathcal{U}} \times \boldsymbol{\mathcal{C}} : x_{ij} > 0\} \cup \{0\}$ is the set of feasible actions (class 0 represents the idling action) when the number of patients of each class in the examination rooms is $\underline{x}$.

The optimal behavior of the physician is an appealing and simple operational rule, supporting implementation in practice.

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

Theorem 1 provides a simple prioritization index for physicians 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 a decision support system) can dynamically estimate the required quantities. The authors have developed a smart phone application that can be used by physicians to facilitate collection of required data and computation of patient priorities.

The above analysis confirms our intuition that a simple priority rule for Phase 2 is optimal. Moreover, the Phase 2 priority rule is consistent with that of Phase 1, since $1/\mu_{ij} = 1/(p_{ij}\hat{\mu}_{ij})$.

6. A Realistic Simulation Analysis of Complexity-Based Triage

Our analytical models of the previous sections suggest that adding patient complexity to the triage process and using appropriate priority rules can improve the ED performance in terms of both patient safety (ROAE) and operational efficiency (LOS). Furthermore, they provide some insights

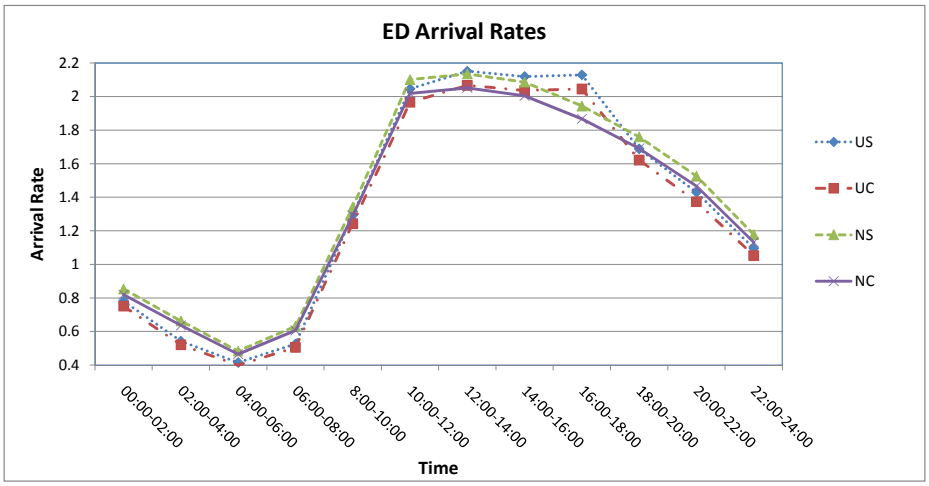


Figure 5 Class dependent arrival rates to the ED for an average day (obtained from a year of data in UMED).

into hospital conditions under which such improvements are more beneficial. In this section, we test the conjectures suggested by our analytic models by means of a detailed ED simulation model. This simulation incorporates many realistic features of the University of Michigan ED (UMED) that are representative of most ED's in large research hospitals, 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. Our base case model uses a year of data from UMED plus time study data from the literature. 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. Following the definition of complex patients in Vance and Spirvulis (2005), we define S patients as those who only require one treatment related interaction and C patients as those requiring two or more treatment related interactions.(To clarify, we do not count social interactions as a treatment related interaction. Furthermore, we would still classify a case as simple if the physician were first to order a test (without spending time treating the patient) and after receiving the results, conduct one treatment visit prior to discharge.) With ESI-4

and 5 patients omitted, we can equate U patients with ESI-2 patients, and N patients with ESI-3 patients for our purposes. Since the majority (about 80% in university of Michigan Hospital ED) of ED patients are composed of ESI-2 and ESI-3 patients, 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. 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) and Vance and Spirvulis (2005), we assume urgency and complexity classifications are subject to 10% and 17% error rates, respectively. For simplicity, 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 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 5. These arrival rates were obtained from a year of UMED data using the ESI levels based on two-hour intervals of the day. However, since patients are not currently triaged based on complexity, we used the empirical results of Vance and Spirvulis (2005) (who found that about 49% of patients are complex) to obtain these arrival rates using a (stationary) splitting mechanism. The resulting pattern illustrated in Figure 5 is similar to those reported in other studies (e.g., Green et al. (2006)). A "thinning" mechanism (see Lewis and Shedler (1979a) and Lewis and Shedler (1979b)) is used to simulate the non-stationary Poisson process arrivals for each class of patients (with rates depicted in Figure 5) in our base case.

Service Process. The ED service process has multiple stages as depicted in the schematic in Figure 4. Each patient goes through one or more phases of 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 4). 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

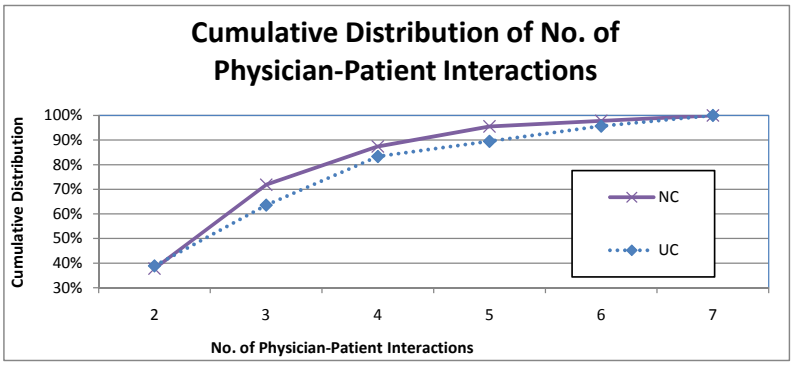


Figure 6 Cumulative number of class-based physician-patient interactions for complex patients (those requiring more than one interaction).

happens after the final visit by the physician and before the patient is discharged home or admitted to the hospital. The duration of each physician interaction is random and its average may depend on the class of the patient and the number of previous interactions. Our data suggest that the first and last interactions are typically longer than the intermediate interactions. As mentioned before and illustrated in Figure 1 (right), S patients are defined to be those who have only one (treatment related) interaction. For C patients, we can estimate the distribution of the number of physician interactions per patient as shown in Figure 6 using data from a detailed time study (see Table 3 of Graff et al. (1993)) (normalized 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.

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 based on a Phase 1 sequencing priority. Phase 1 is usually performed by a charge nurse. 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 will depend on the type of triage being used. For urgency-based triage, we assume U patients get priority over N patients in both Phases 1 and 2. For complexity-based triage, patients are prioritized in both Phases according to the strict 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 (see Proposition 2).

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 more than 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. We then perform a sensitivity analysis to understand the effect of number of both beds and physicians on the benefit of complexity-based triage. For simplicity, we do not consider end of shift effects and/or variations in the level of staff available. Furthermore, we consider test facilities (ancillary services) as exogenous resources (i.e., test times are independent of the volume of ED patients) because these facilities often handle many other patients besides those from the ED.

Adverse Events. Adverse events are simulated using Poisson processes with rates that depend on the class of patients, as well as the phase of service. Specifically, we assume that U patients have a higher rate of adverse events than N patients, and that after patients enter an exam room (Phase 2 of service), their rate of adverse events decreases by 60% (in our base case) relative to their rates in the waiting area (Phase 1 of service). 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 a C++ framework and makes use of a cyclo-stationary model 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 (which was observed to be sufficient because correlations in the ED flow are very small for spans of two or more days). The number of replications (5000) was chosen so that the confidence intervals are tight enough that (1) the sample averages are reliable, and (2) we can omit 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-based triage over urgency-based triage.

6.1. Performance of Complexity-Based 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 Phases 2. This leads to the following:

Observation 1. *In the base case, implementing complexity-based triage improves ROAE and LOS by* 9.41% *(0.16 events/hr) and* 7.68% *(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-based 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 7.95% and 7.01% in ROAE and LOS, respectively. Hence, it appears that the benefits of complexity-based triage are 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.

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 a considerable amounts of time while being tested and waiting for test results, and (b) each physicians is handling a limited number of patients at a time (with a constrained 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-based triage systems, then complexity-based triage achieves improvements of 8.58% and 6.15% 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-based triage system prioritizes patients in Phase 2 by complexity and urgency ($US > UC > NS > NC$), then complexity-based triage achieves improvements of 13.09% and 9.11% 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.*

6.2. How to Define Complex Patients?

In the previous section, we investigated the benefit of complexity-based triage using the approach of Vance and Spirvulis (2005) to define complex patients as those requiring at least two (treatment related) interactions with a physician. This results in a nearly even split between complex and simple patients (49% C vs. 51% S), as well as substantial heterogeneity between their treatment time (both of which were predicted in Proposition 3 to be factors that improve the performance of complexity-based triage). But we could use other standards for defining a patient to be complex. In Figure 7, we illustrate the impact of complexity-based triage on ROAE and LOS when complex patients are defined to be as those with more than one (resulting in 49% C patients), more than two (resulting in 39% C patients), and more than three (resulting in 30% C patients) interactions. From this we conclude:

Observation 3. *If the number of (treatment related) interactions is used as the metric for patient complexity, the benefit of complexity-based 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-based triage declines.

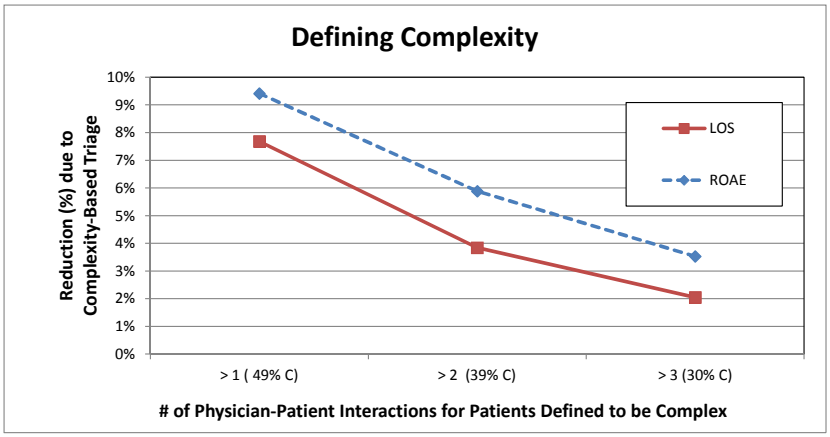


Figure 7 Performance of complexity-based triage when defining complex patients to be those having more than one, more than two, and more than three physician-patient interactions.

6.3. The Effect of ED Resource Levels

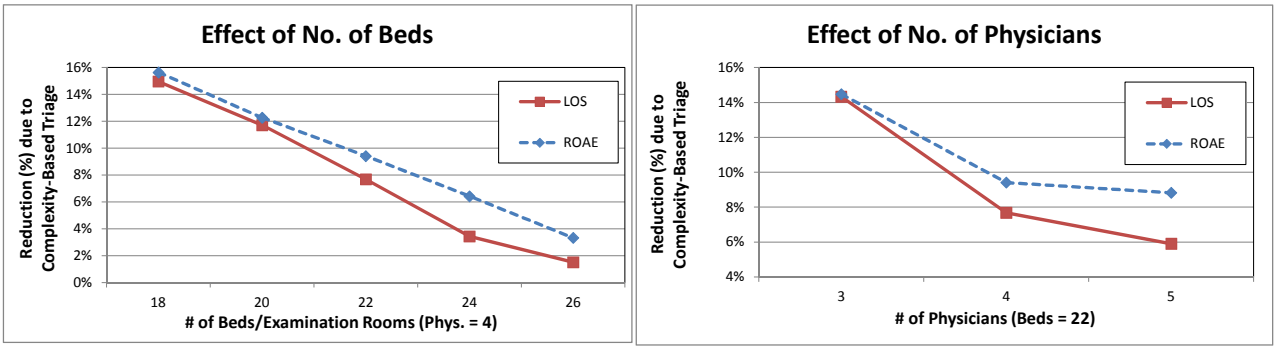
Another factor predicted by Proposition 3 to favor complexity-based triage is resource utilization. In that proposition, resources refer to physicians and examination rooms (which are indistinguishable in the single-stage simplified ED model). Hence, we expect higher utilization of either physicians or examination rooms to increase the benefit of complexity-based triage. Figure 8 illustrates the percentage improvement (in terms of ROAE and LOS) of complexity-based triage over urgency-based triage for varying numbers of examination rooms and physicians. In addition to the LOS for patient classes considered (i.e., ESI 2 and 3) with 4 physicians, this figure also presents the average time spent in Phase 1, labeled as Time to Room (TTR), under each triage system. From this figure we observe the following:

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

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

6.4. The Effect of Misclassification

Finally, we investigate the impact of complexity-based misclassification errors, which are inevitable in any triage system. Figure 9 (left) shows the benefits (in ROAE and LOS) of complexity-based



Triage Policy	Phys. = 4, Beds = 18		Phys. = 4, Beds = 20		Phys. = 4, Beds = 22		Phys. = 4, Beds = 24		Phys. = 4, Beds = 26	
	TTR (hrs)	LOS (hrs)	TTR (hrs)	LOS (hrs)	TTR (hrs)	LOS (hrs)	TTR (hrs)	LOS (hrs)	TTR (hrs)	LOS (hrs)
Urgency-Based	31.57	35.24	10.66	14.43	3.91	7.81	1.82	5.81	1.16	5.25
Complexity-Based	26.26	29.97	8.95	12.74	3.3	7.21	1.59	5.61	1.04	5.17

Figure 8 The effect of resources (beds and physicians) on the benefit of complexity-based triage over the current practice of urgency-based triage [Left: the effect of beds (4 physicians); Right: the effect of physicians (22 beds)].

triage over urgency-based triage for variations of the base case, in which complexity misclassification error rates range from 5% to 25%. Figure 9 (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 9 (right) considers asymmetric error rates while keeping the average misclassification rate constant and equal to the base-case value of 17% (reported in the empirical study of ((Vance and Spirvulis (2005))). From these figures, we observe the following:

Observation 5. *The benefit of complexity-based triage is 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.

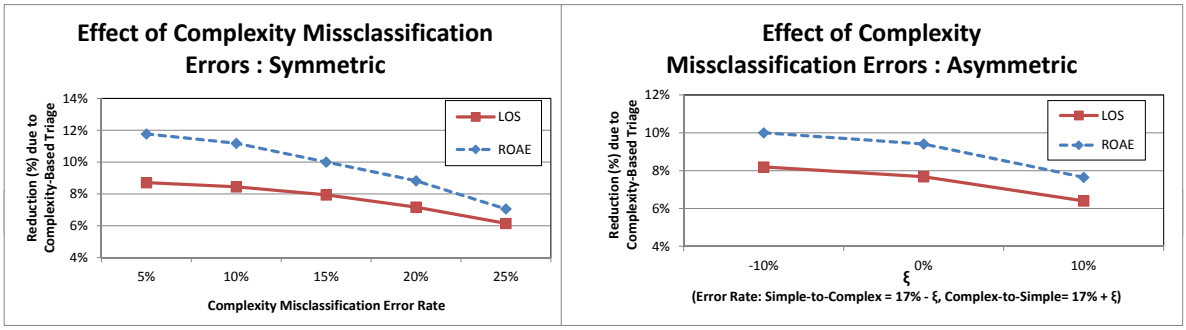


Figure 9 The effect of complexity misclassification error rates on the benefit of a complexity-based triage (compared to an urgency-based only) [Left: symmetric misclassification; Right: asymmetric misclassification].)

In contrast, a simple-to-complex misclassification error moves a 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.

6.5. Complexity-Based Streaming

In the previous sections, we investigated the benefit of collecting and using complexity-based information to prioritize patients in the ED. But another way to make use of this information is to separate patients into different service streams for simple and complex patients (somewhat analogous to the admit/discharge streaming implemented in Flinders Medical Center (King et al. (2006)) with complexity information used in place of admit/discharge predictions). We are interested in whether such streaming is more effective than pooling-based prioritization.

To investigate this design question we raised in the Introduction, we examine a *complexity-based streaming* patient flow design in which two service streams of patients are created: 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 overcomes the "anti-pooling" disadvantage of streaming, we allow the resources to be assigned to the other stream as needed, which is a feature we found to be useful in Saghafian et al. (2010). For instance, 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. In this design, we assume that patients in each stream and in both Phases 1 and 2 are prioritized according to their ESI level.

Since simple and complex patients are separated, lean process improvement techniques can be implemented to improve and standardize service, particularly on the simple side for which the repetitive treatment processes can be organized in a clear, flow-shop like path. In Figure 10, we compare the performance of the complexity-based streaming design, with and without such lean improvements, against that of urgency-based pooling (current practice) and complexity-based pooling (i.e., a pooling design where Phase 1 and Phase 2 are based on the optimal priority rule using complexity-based triage information). The system with lean implementation assumes the service

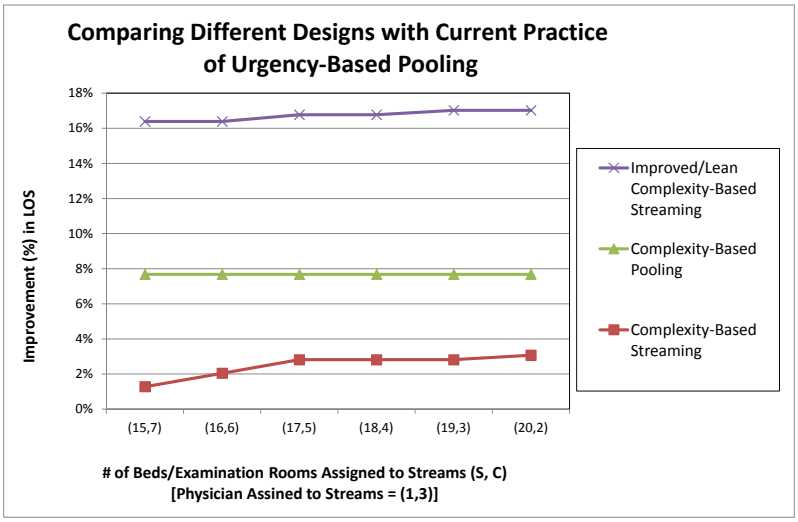


Figure 10 Performance of different patient flow designs compared to the current practice (Urgency-Based Pooling).

rate for each interaction with the simple patients improves by 10%; however, no change occurs for complex patients. This is a conservative estimate of the impact of a lean transformation. Note that the streaming layout facilitates this improvement by grouping simple tasks in a single line. Under pooled designs the mixture of simple and complex patients makes a smooth efficient flow extremely difficult. It should be also noted that some lean improvements may be possible for complex stream, but we conservatively ignore that here. Figure 10 compares performance in terms of LOS, but we have observed similar results for the ROAE criterion. These results lead to the following:

Observation 6. *Without lean improvements, complexity-based streaming is still better than current pooling practice, but worse than complexity-based pooling. With lean improvements (made only to the simple stream), complexity-based streaming can achieve a substantial advantage over complexity-based pooling.*

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 the triage system currently in use in practice, complexity-based 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 also investigate effective but implementable policies for prioritizing patients in the ED. We show that the current practice of prioritizing patients purely based on urgency (e.g., ESI 2 over 3 in the main ED) is suboptimal, and it is essential to take into account a measure of patient complexity. This can address many of the performance limitations of the current triage system in ED's that are widely reported by clinicians (see, e.g., Welch and Davidson (2011) and the references therein).

We find that a simple and fast classification scheme, which defines patients to be simple if they require only a single treatment related interaction (and complex otherwise) works very well as the basis for complexity-based triage as 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. This classification scheme has been empirically shown (see Vance and Spirvulis (2005)) to be feasible for nurses to implement at triage with reasonable accuracy, and hence, appears to be a promising enhancement of the triage process.

To accomplish this research, we developed new models, contributed several analytical contributions, collected hospital data, and developed high-fidelity simulations. We advanced the analysis of priority queueing systems under misclassification errors as well as continuous-time MDP analysis with unbounded transition rates, for which the traditional method of uniformization fails. Using these technical innovations, we show that new, extended versions of the $c\mu$ rule can provide effective guidelines for prioritizing patients in both Phase 1 and Phase 2 of service in the ED, even when many practical conditions in the ED are considered.

Our analyses indicate that complexity-based triage can yield substantial safety and efficiency improvements even if complexity information is only used to prioritize patients up to the point where they enter 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-based 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-based triage is an effective way for ED's to improve safety and reduce congestion without adding expensive human or physical capacity.

We further investigate a new patient flow design, in which complexity-based triage information is used to separate simple and complex patients into two streams. Our results show that, when combined with improvements achieved through implementation of lean methods on the "simple" patient service stream, this complexity-based streaming design can take advantage of complexity-based triage information to achieve even greater gains.

References

American Hospital Association. 2002. Emergency Department overload: a growing crisis. The results of the American Hospital Association survey of Emergency Department (ED) and hospital capacity. Falls Church, VA: American Hospital Association, 2002.

Argon, N.T., S. Ziya. 2009. Priority assignment under imperfect information on customer type identities. *Manufacturing Service Oper. Management* **11**(4) 674–693.

Buyukkoc, C., P. Varaiya, J. Walrand. 1985. The $c\mu$ rule revisited. *Adv. Appl. Prob.* **17** 237–238.

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.

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.

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.

Gordon, J.A., J. Billings, B.R. Asplin et al. 2001. Safety net research in emergency medicine: proceedings of the academic emergency medicine consensus conference on the unraveling safety net. *Acad. Emerg. Med.* **8** 10249.

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

Guo, X., K. Liu. 2001. A note on optimality conditions for continuous-time Markov Decision Processes with aevrage cost criterion. *IEEE Trans. on Aut. Contr.* **46**(12) 1984–1989.

Han, J.H., D.J. France, S.R. Levin et. al. 2010. The effect of physician triage on Emergency Department length of stay. *J. of Emerg. Med.* **39**(2) 227–233.

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.

Holland, L.L., L.L. Smith, K.E. Blick. 2005. Reducing laboratory turnaround time outliers can reduce Emergency Department patient length of stay. *Amer. J. Clin. Path.* **124** 672–674.

Ierson, K.V., J.C. Moskop. 2007. Triage in medicine, part I: Concept, history, and types. *Annals of Emerg. Med.* **49**(3) 275–281.

Institute of Medicine. 2007. *Hospital-Based Emergency Care: At the Breaking Point*. National Academies Press, London.

Jaiswal, N.K. 1968. *Priority Queues*. Academic Press, New York, New York.

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.

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.

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.

Moskop, J.C., K.V. Ierson. 2007. Triage in medicine, part II: Underlying values and principles. *Annals of Emreg. Med.* **49**(3) 282–287.

Russ, S., I. Jones, D. Aronsky et. al. 2010. Placing physician orders at triage: The effect on length of stay. *Annals of Emerg. Med.* **56**(1) 27–33.

Saghafian, S., W.J. Hopp, M.P. Van Oyen, J.S. Desmond, S.L. Kronick. 2010. Patient streaming as a mechanism for improving responsiveness in Emergency Departments. Working paper, Dept. of IOE, Univ. of Michigan.

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.

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.

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.

Steindel, S.J., P.J. Howanitz. 1997. Changes in Emergency Department turnaround time performance from 1990 to 1993. *Arch. Path. & Lab. Med.* **121** 1031–1041.

Thomas, E.J., D.M. Studdert, H.R. Burstin et. al. 2000. Incidene and type of adverse events and negligent care in Utah and Colorado. *Medical Care* **38**(3) 261–271.

Trzeciak, S., E.P. Rivers. 2003. Emergency Department overcrowding in the United States: an emerging threat to patient safety and public health. *Emerg. Med. J.* **20**(5) 402–405.

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.

Veatch, Michael H. 2010. A $c\mu$ rule for prallel servers with two tiered $c\mu$ preference. Working Paper, Math. Dept., Gordon College.

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.

Wolf, R.W. 1989. *Stochastic Modeling and the Theory of Queues*. Prentice Hall, Englewood Cliffs, New Jersey.

Online Appendix (Proofs)

Proof of Proposition 1: The proof of part (i) follows directly from comparing (4) and (5). 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 $\theta'_U \mu'_U - \theta'_N \mu'_N = \big[\lambda_N \lambda_U \mu_N \mu_U (\theta_U \mu_U - \theta_N \mu_N)(1 - \gamma_N - \gamma_U)\big] / \big[\big(\lambda_N \mu_U \gamma_N + \lambda_U \mu_N (1 - \gamma_U)\big)\big(\lambda_N \mu_U (1 - \gamma_N) + \lambda_U \mu_N \gamma_U\big)\big]$. Combining these two results completes the proof of part (ii), as the sign of the numerator changes when the sum of errors exceeds 1. □

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

$$R_\pi^{\mathcal{U}\cup\mathcal{C}} = \sum_{i\in\mathcal{U}} \sum_{j\in\mathcal{C}} \theta_{ij}\lambda_{ij} W_{ij}^\pi, \tag{EC.1}$$

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 (1) and (EC.1), when $\theta_{UC}\mu_{UC} \geq \theta_{NS}\mu_{NS}$, we have:

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

where the inequality follows from (2) and (3) 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:

$$\begin{aligned} R_*^{\boldsymbol{\mathcal{U}}\cup\boldsymbol{\mathcal{C}}} &= \lambda\mathbb{E}(s^2)\Big[\frac{\lambda_{US}\theta_{US}}{2(1-\rho_{US})} + \frac{\lambda_{NS}\theta_{NS}}{2(1-\rho_{US})(1-\rho_{US}-\rho_{NS})} + \frac{\lambda_{UC}\theta_{UC}}{2(1-\rho_{US}-\rho_{NS})(1-\rho_{US}-\rho_{NS}-\rho_{UC})} \\ &\qquad + \frac{\lambda_{NC}\theta_{NC}}{2(1-\rho_{US}-\rho_{UC}-\rho_{NS})(1-\rho_{US}-\rho_{UC}-\rho_{NS}-\rho_{NC})}\Big], \qquad \text{(EC.3)} \end{aligned}$$

and similar to the previous case, it can be easily seen that $R_*^{\boldsymbol{\mathcal{U}}\cup\boldsymbol{\mathcal{C}}} \leq R_*^{\boldsymbol{\mathcal{U}}}$. 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). □

LEMMA EC.2 **(Perfect Classification - Attractiveness)**. *In the simplified single-stage ED model, perfect complexity-based 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 EC.2: To show the result, first consider the case where under the $\boldsymbol{\mathcal{U}}\cup\boldsymbol{\mathcal{C}}$ classification it is optimal to follow the priority order US, UC, NS, NC, and under the $\boldsymbol{\mathcal{U}}$ classification, it is

optimal to follow the priority order U, N (i.e., prioritizing urgent patients first). Let $f = R_*^{\boldsymbol{\mathcal{U}}\cup\boldsymbol{\mathcal{C}}} - R_*^{\boldsymbol{\mathcal{U}}}$, and notice that with $\mu_{iC} = \mu_C$ and $\mu_{iS} = \mu_S$ $(\forall i \in \boldsymbol{\mathcal{U}})$, and $\theta_{Uj} = \theta_U$ and $\theta_{Nj} = \theta_N$ $(\forall j \in \boldsymbol{\mathcal{C}})$ (i.e., when complexity is based only on set $\boldsymbol{\mathcal{C}}$ and urgency is based only on set $\boldsymbol{\mathcal{U}}$), from (EC.2) and (2) we have:

$$f = -[\frac{\theta_U \lambda_{US}\lambda_{UC}(1/\mu_C - 1/\mu_S)}{2(1-\rho_U)} + \frac{\theta_N \lambda_{NC}\lambda_{NS}(1/\mu_C - 1/\mu_S)}{2(1-\rho_U)(1-\rho)}]. \tag{EC.4}$$

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 (EC.4), 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 (EC.2) or (EC.3), and either (2) or (3), depending on the optimal priority rule under $\boldsymbol{\mathcal{U}} \cup \boldsymbol{\mathcal{C}}$ and $\boldsymbol{\mathcal{U}}$ classifications, respectively. □

Proof of Proposition 2: The proof of part (i) follows directly from the proof of part (i) of Lemma EC.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 EC.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 rate of adverse events, and all the other rates with their error impacted counterparts. □

Proof of Proposition 3: The proof of parts (i) - (iii) follows mainly from the proof of Lemma EC.2 (found earlier in this appendix). First, consider the case where under the $\boldsymbol{\mathcal{U}}' \cup \boldsymbol{\mathcal{C}}'$ (i.e., imperfect urgency and complexity) classification it is optimal to follow the priority order US, UC, NS, NC, and under the $\boldsymbol{\mathcal{U}}'$ (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 $f = R_*^{\boldsymbol{\mathcal{U}}' \cup \boldsymbol{\mathcal{C}}'} - R_*^{\boldsymbol{\mathcal{U}}'}$, and after replacing rates with their error impacted counterparts in (EC.4) we have:

$$f = -[\frac{\theta'_U \lambda'_{US} \lambda'_{UC} (1/\mu'_C - 1/\mu'_S)}{2(1-\rho'_U)} + \frac{\theta'_N \lambda'_{NC} \lambda'_{NS} (1/\mu'_C - 1/\mu'_S)}{2(1-\rho'_U)(1-\rho')}]. \tag{EC.5}$$

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 EC.2, it can bee 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 $(\underline{1}/\underline{\mu'})^T = (A\,(\underline{\lambda}/\underline{\mu})^T)/\underline{\lambda'}$, where A is defined in (6). Thus, similar to the proof of part (ii) of Lemma EC.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 EC.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. □

Proof of Theorem 1: To show the result, we use an *interchange* argument; we show that if classes $uc \in \boldsymbol{\mathcal{U}} \times \boldsymbol{\mathcal{C}}$ and $sl \in \boldsymbol{\mathcal{U}} \times \boldsymbol{\mathcal{C}}$ 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 $n \in 1, 2, ..., N$. (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 $n \to \infty$, 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 $k \to \infty$, 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 (7). With β denoting the discount factor, the optimal discounted cost when there are $n+1$ (uniformized) periods to go is

$$\begin{aligned} V_{n+1}^k(\underline{x},\underline{y}) = \frac{1}{\psi_k}\Big[\hat{\underline{\theta}}\,(\underline{x}+\underline{y})^T + \beta\big[\sum_{ij\in\boldsymbol{\mathcal{U}}\times\boldsymbol{\mathcal{C}}} [\lambda_{ij} V_n^k(\underline{x}+\underline{e}_{ij},\underline{y}) + (y_{ij}\wedge k)\,\eta\, V_n^k(\underline{x}+\underline{e}_{ij},\underline{y}-\underline{e}_{ij})] \\ + \min_{a\in\mathcal{A}(\underline{x})}\Big\{\sum_{ij\in\boldsymbol{\mathcal{U}}\times\boldsymbol{\mathcal{C}}} \mathbb{1}_{\{a=ij\}}\hat{\mu}_{ij}\,[p_{ij}\,V_n^k(\underline{x}-\underline{e}_{ij},\underline{y}) + (1-p_{ij})V_n^k(\underline{x}-\underline{e}_{ij},\underline{y}+\underline{e}_{ij})] \\ + \big(\psi_k - \sum_{ij\in\boldsymbol{\mathcal{U}}\times\boldsymbol{\mathcal{C}}} [\lambda_{ij} + (y_{ij}\wedge k)\,\eta + \mathbb{1}_{\{a=ij\}}\hat{\mu}_{ij}]\big)\,V_n^k(\underline{x},\underline{y})\Big\}\big]\Big], \end{aligned} \tag{EC.6}$$

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

$$\begin{aligned} V_{n+1}^k(\underline{x},\underline{y}) = \frac{1}{\psi_k}\Big[\hat{\underline{\theta}}\,(\underline{x}+\underline{y})^T + \beta\big[\sum_{ij\in\boldsymbol{\mathcal{U}}\times\boldsymbol{\mathcal{C}}} [\lambda_{ij} V_n^k(\underline{x}+\underline{e}_{ij},\underline{y}) + (y_{ij}\wedge k)\,\eta\, V_n^k(\underline{x}+\underline{e}_{ij},\underline{y}-\underline{e}_{ij})] \\ - \max_{a\in\mathcal{A}(\underline{x})}\Big\{\sum_{ij\in\boldsymbol{\mathcal{U}}\times\boldsymbol{\mathcal{C}}} \mathbb{1}_{\{a=ij\}}\hat{\mu}_{ij}\,[p_{ij}\Delta_{ij}^{y}\,V_n^k(\underline{x}-\underline{e}_{ij},\underline{y}) + \Delta_{ij}^{x,y}V_n^k(\underline{x}-\underline{e}_{ij},\underline{y}+\underline{e}_{ij})] \\ + \big(\psi_k - \sum_{ij\in\boldsymbol{\mathcal{U}}\times\boldsymbol{\mathcal{C}}} [\lambda_{ij} + (y_{ij}\wedge k)\,\eta]\big)\,V_n^k(\underline{x},\underline{y})\Big\}\big]\Big], \end{aligned} \tag{EC.7}$$

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 (EC.7), 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 $x_{uc}, x_{sl} > 0$:

$$\begin{aligned} \hat{\mu}_{uc}\,[p_{uc}\Delta_{uc}^{y}\,V_n^{k,\pi}(\underline{x}-\underline{e}_{uc},\underline{y}) + \Delta_{uc}^{x,y}V_n^{k,\pi}(\underline{x}-\underline{e}_{uc},\underline{y}+\underline{e}_{uc})] \\ \geq \hat{\mu}_{sl}\,[p_{sl}\Delta_{sl}^{y}\,V_n^{k,\hat{\pi}}(\underline{x}-\underline{e}_{sl},\underline{y}) + \Delta_{sl}^{x,y}V_n^{k,\hat{\pi}}(\underline{x}-\underline{e}_{sl},\underline{y}+\underline{e}_{sl})]. \end{aligned} \tag{EC.8}$$

To show property (EC.8), 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 $V^{k,\pi}_{n+1}(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc})$ and $V^{k,\pi}_{n+1}(\underline{x} - \underline{e}_{uc}, \underline{y})$ using (EC.7), and subtracting the results we have $\Delta^{y}_{uc} V^{k,\pi}_{n+1}(\underline{x} - \underline{e}_{uc}, \underline{y}) =$

$$\begin{aligned}
\frac{1}{\psi_k}\Big[\hat{\theta}_{uc} + \beta\big[\sum_{ij \in \boldsymbol{\mathcal{U}} \times \boldsymbol{\mathcal{C}}} [\lambda_{ij} \Delta^{y}_{uc} V^{k,\pi}_{n}(\underline{x}, \underline{y}) + (y_{ij} \wedge k)\, \eta\, \Delta^{y}_{uc} V^{k,\pi}_{n}(\underline{x} + \underline{e}_{ij} - \underline{e}_{uc}, \underline{y} - \underline{e}_{ij})] \\
+ \mathbb{1}_{\{y_{uc} < k\}}\, \eta\, V^{k,\pi}_{n}(\underline{x}, \underline{y} - \underline{e}_{uc}) \\
- \hat{\mu}_{uc}\, [p_{uc} \Delta^{y}_{uc} \Delta^{y}_{uc} V^{k,\pi}_{n}(\underline{x} - 2\underline{e}_{uc}, \underline{y}) + \Delta^{y}_{uc} \Delta^{x,y}_{uc} V^{k,\pi}_{n}(\underline{x} - 2\underline{e}_{uc}, \underline{y} + \underline{e}_{ij})] \\
+ \big(\psi_k - \sum_{ij \in \boldsymbol{\mathcal{U}} \times \boldsymbol{\mathcal{C}}} [\lambda_{ij} + (y_{ij} \wedge k)\, \eta]\big)\, \Delta^{y}_{uc} V^{k}_{n}(\underline{x} - \underline{e}_{uc}, \underline{y}) - \mathbb{1}_{\{y_{uc} < k\}}\, \eta\, V^{k,\pi}_{n}(\underline{x} - \underline{e}_{uc}, \underline{y})\big]\Big].
\end{aligned} \tag{EC.9}$$

Similarly, we can derive $\Delta^{x,y}_{uc} V^{k,\pi}_{n+1}(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc})$ using (EC.7) 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 $\Delta^{x,y}_{uc} V^{k,\pi}_{n+1}(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc}) =$

$$\begin{aligned}
\frac{1}{\psi_k}\Big[\beta\big[\sum_{ij \in \boldsymbol{\mathcal{U}} \times \boldsymbol{\mathcal{C}}} [\lambda_{ij} \Delta^{x,y}_{uc} V^{k,\pi}_{n}(\underline{x}, \underline{y} + \underline{e}_{uc}) + (y^{+}_{ij} \wedge k)\, \eta\, \Delta^{x,y}_{uc} V^{k,\pi}_{n}(\underline{x} + \underline{e}_{ij} - \underline{e}_{uc}, \underline{y} - \underline{e}_{ij} + \underline{e}_{uc})] \\
- \mathbb{1}_{\{y_{uc} < k\}}\, \eta\, V^{k,\pi}_{n}(\underline{x}, \underline{y}) \\
- \hat{\mu}_{uc}\, [p_{uc} \Delta^{x,y}_{uc} \Delta^{y}_{uc} V^{k,\pi}_{n}(\underline{x} - 2\underline{e}_{uc}, \underline{y} + \underline{e}_{uc}) + \Delta^{x,y}_{uc} \Delta^{x,y}_{uc} V^{k,\pi}_{n}(\underline{x} - 2\underline{e}_{uc}, \underline{y} + \underline{e}_{ij} + \underline{e}_{uc})] \\
+ \big(\psi_k - \sum_{ij \in \boldsymbol{\mathcal{U}} \times \boldsymbol{\mathcal{C}}} [\lambda_{ij} + (y^{+}_{ij} \wedge k)\, \eta]\big)\, \Delta^{x,y}_{uc} V^{k}_{n}(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc}) + \mathbb{1}_{\{y_{uc} < k\}}\, \eta\, V^{k,\pi}_{n}(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc})\big]\Big],
\end{aligned} \tag{EC.10}$$

where $y^{+}_{ij} = y_{ij}$ for all $ij \neq uc \in \boldsymbol{\mathcal{U}} \times \boldsymbol{\mathcal{C}}$, and $y^{+}_{uc} = y_{uc} + 1$. In a similar way, and by using action $a = sl$ (policy $\hat{\pi}$) in (EC.7) quantities $\Delta^{y}_{sl} V^{k,\hat{\pi}}_{n+1}(\underline{x} - \underline{e}_{sl}, \underline{y})$ and $\Delta^{x,y}_{sl} V^{k,\hat{\pi}}_{n+1}(\underline{x} - \underline{e}_{sl}, \underline{y} + \underline{e}_{sl})$ can be computed. Next, to check property (EC.8) for $n+1$, multiply (EC.9) by $p_{uc}\, \hat{\mu}_{uc}$, and (EC.10) by $\hat{\mu}_{uc}$ and add up the results. Similarly, multiply $\Delta^{y}_{sl} V^{k,\hat{\pi}}_{n+1}(\underline{x} - \underline{e}_{sl}, \underline{y})$ and $\Delta^{x,y}_{sl} V^{k,\hat{\pi}}_{n+1}(\underline{x} - \underline{e}_{sl}, \underline{y} + \underline{e}_{sl})$ 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

$$\begin{aligned}
\hat{\mu}_{uc}\, [p_{uc} \Delta^{y}_{uc} V^{k,\pi}_{n+1}(\underline{x} - \underline{e}_{uc}, \underline{y}) + \Delta^{x,y}_{uc} V^{k,\pi}_{n+1}(\underline{x} - \underline{e}_{uc}, \underline{y} + \underline{e}_{uc})] \\
- \hat{\mu}_{sl}\, [p_{sl} \Delta^{y}_{sl} V^{k,\hat{\pi}}_{n+1}(\underline{x} - \underline{e}_{sl}, \underline{y}) + \Delta^{x,y}_{sl} V^{k,\hat{\pi}}_{n+1}(\underline{x} - \underline{e}_{sl}, \underline{y} + \underline{e}_{sl})] \geq 0,
\end{aligned} \tag{EC.11}$$

which establishes property (EC.8) 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, a non-idling strict priority rule is optimal for all n. Next, taking the limit as $n \to \infty$ 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). □