This article was downloaded by: [129.219.247.33] On: 16 August 2014, At: 09:39
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA

Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

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

Soroush Saghafian, Wallace J. Hopp, Mark P. Van Oyen, Jeffrey S. Desmond, Steven L. Kronick

To cite this article:
Soroush Saghafian, Wallace J. Hopp, Mark P. Van Oyen, Jeffrey S. Desmond, Steven L. Kronick (2014) Complexity-Augmented Triage: A Tool for Improving Patient Safety and Operational Efficiency. Manufacturing & Service Operations Management 16(3):329-345. http://dx.doi.org/10.1287/msom.2014.0487

Full terms and conditions of use: http://pubsonline.informs.org/page/terms-and-conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2014, INFORMS

Please scroll down for article—it is on subsequent pages

INFORMS is the largest professional society in the world for professionals in the fields of operations research, management science, and analytics.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

RIGHTSLINK

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 16, No. 3, Summer 2014, pp. 329–345
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.2014.0487
© 2014 INFORMS

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

Soroush Saghafian
Industrial Engineering, School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, Arizona 85281, soroush.saghafian@asu.edu

Wallace J. Hopp
Ross School of Business, University of Michigan, Ann Arbor, Michigan 48109, whopp@umich.edu

Mark P. Van Oyen
Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan 48109, vanoyen@umich.edu

Jeffrey S. Desmond, Steven L. Kronick
Department of Emergency Medicine, University of Michigan Health System, Ann Arbor, Michigan 48109 {jsdesmo@med.umich.edu, skronick@med.umich.edu}

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

Keywords: healthcare operations; emergency department; triage; priority queues; patient prioritization; Markov decision processes
History: Received: October 30, 2012; accepted: January 30, 2014. Published online in *Articles in Advance* May 9, 2014.

1. Introduction

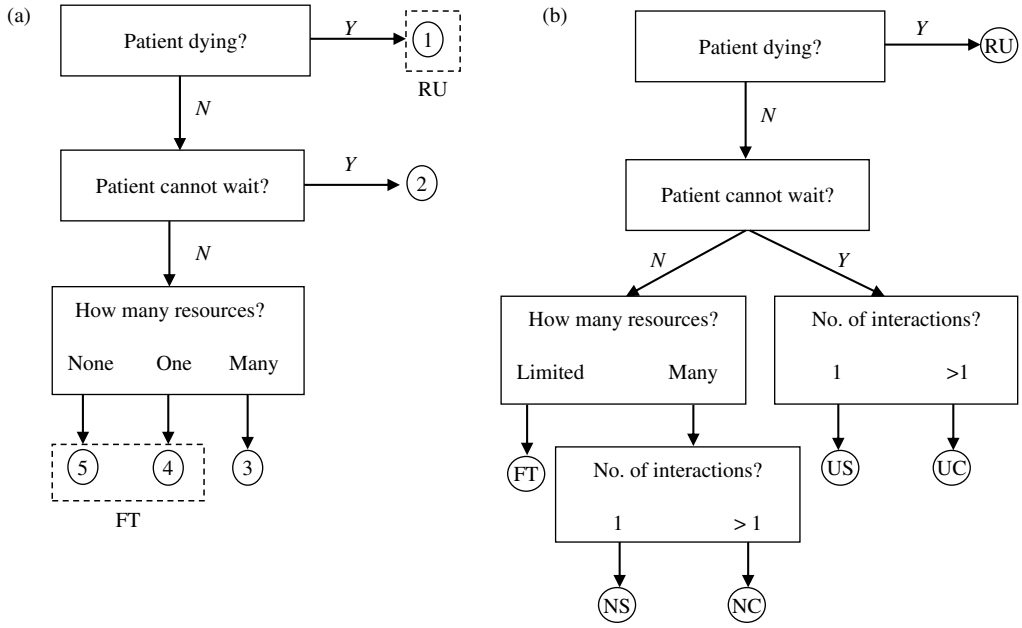
Triage (a word derived from the French verb "trier," meaning "to sort") refers to the process of sorting and prioritizing patients for care. FitzGerald et al. (2010, p. 80) noted that there are two main purposes for triage: "[1] to ensure that the patient receives the level and quality of care appropriate to clinical need (clinical justice), and [2] that departmental resources are most usefully applied (efficiency) to this end." (See Moskop and Ierson 2007 for further discussion of the underlying principles and goals of triage.)

Most triage systems used around the world consider only urgency and so only address the clinical justice purpose of triage. For instance, the Australasian Triage Scale (ATS), the Manchester Triage Scale (MTS), and the Canadian Triage Acuity Scale (CTAS) differ in their details, but all classify patients strictly in terms of urgency. In the United States, many hospital emergency departments (EDs) continue to use a traditional urgency-based three-level triage scale, which categorizes patients into emergent, urgent, and nonurgent classes. But a growing majority of U.S. hospitals have adopted five-level triage systems (see Fernandes et al. 2005), which seek to address efficiency by incorporating an estimate of resources (e.g., tests) required. In these systems (a typical version of which is illustrated in Figure 1(a)), urgent patients who cannot wait are classified as level 1 or 2, whereas nonurgent patients who can wait are classified as level 3, 4, or 5. Level 4 and level 5 patients are usually directed to a fast track (FT) area, whereas level 1 patients are almost always moved immediately to a resuscitation unit (RU). Level 2 and level 3 patients, who represent the majority of patients at large academic hospitals (about 80% at the University of Michigan Health System ED (UMHSED)), are served in the main area of the ED with priority given to level 2 patients. Since five-level systems do not differentiate between level 2 and level 3 patients in terms of complexity,

Downloaded from informs.org by [129.219.247.33] on 16 August 2014, at 09:39 . For personal use only, all rights reserved.

RIGHTSLINK

Figure 1 (a) Typical Five-Level Triage System (see, e.g., Gilboy et al. 2005); (b) Proposed Complexity-Augmented Triage System



Note. RU, resuscitation unit; FT, fast track; NS, nonurgent simple; NC, nonurgent complex; US, urgent simple; UC, urgent complex.

patients in the main ED (about 80% of patients) are still sorted and prioritized purely on the basis of urgency. Hence, although five-level triage systems represent an improvement over traditional three-level triage scales, they remain urgency-based systems for the majority of patients.

In this paper, we propose an augmented triage system, which we term *complexity-augmented* triage, that can significantly improve performance of the main ED with respect to both clinical justice and efficiency. This poses two challenges: (a) deciding what information to collect at triage, and (b) determining how to use the information to improve performance. There are two main choices for the latter: prioritization and streaming. But they can be combined by using some information to separate patients into streams and some other information to prioritize them within the streams. This poses an additional question: what information to use to stream patients and what information to use to prioritize them?

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

Referring to procedures, investigations, or consultations as "interactions," we propose the new complexity-augmented triage process depicted in Figure 1(b). Unlike a conventional five-level system that makes no complexity distinction among the levels 2 and 3 patients that make up the majority of ED patients, our proposed system systematically classifies them in terms of complexity. The additional step required in triage (i.e., predicting whether the patient will need two or more interactions) can be performed in seconds, and hence, does not add any significant amount of time to triage. However, it is unclear how much this additional information can improve the ED performance in terms of risk of adverse events (clinical justice) and average length of stay (efficiency), since it is subject to misclassification errors. To clarify this and related issues, we make use of a combination of analytic and simulation models calibrated with hospital data to address the following:

1. *Prioritization*: How should EDs use complexity-augmented triage information to prioritize patients?

Downloaded from informs.org by [129.219.247.33] on 16 August 2014, at 09:39 . For personal use only, all rights reserved.

RIGHTSLINK

Downloaded from informs.org by [129.219.247.33] on 16 August 2014, at 09:39 . For personal use only, all rights reserved.

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

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

4. (*Patient Flow*) *Design*: Is complexity information more effective if used to prioritize patients or to separate patients into streams?

The main contribution of this paper is to provide insights of value to ED managers by addressing the above questions. However, the above questions also require addressing some technical challenges: (1) In the ED, upfront triage misclassifications are inevitable. We incorporate misclassifications through a linear transformation of control indices so that they represent "error-impacted" rates, which use only information from historical data. This leads to a modified version of the well-known $c\mu$ rule for the case with customer misclassification (in which control indices are replaced with their linearly transformed "error-impacted" counterparts). Furthermore, although these results are obtained by modeling the occurrence of adverse events as Poisson processes, in Online Appendix C,[1] we use sample path arguments and appropriate notions of stochastic ordering to demonstrate the robustness of the priority rules under more general adverse event occurrence processes. (2) To provide guidance for ED physicians on how to prioritize patients within the examination rooms (when they have a choice of what patient to see next), we develop a Markov decision process (MDP) model. A challenging feature of this model, which is common in many other health delivery settings, is that patients are sent for tests (e.g., MRI, CT scan, x-ray, etc.) and are unavailable to the physician during testing. In such a setting, the physician must consider both the current and the future availability of the patients when making decisions. This type of problem usually results in complex state-dependent optimal control policies. However, we show how a simple-to-implement rule that relies only on historical data defines the optimal policy for ED physicians. (3) Because of unbounded transition rates, the continuous MDP model of patient prioritization within examination rooms cannot use the conventional method of uniformization of Lippman (1975). We contribute by showing how one can use a sequence of MDPs, each with bounded transition rates, to derive an optimal policy for the original MDP.

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

[1] The online appendices are available as supplemental material at http://dx.doi.org/10.1287/msom.2014.0487.

2. Literature Review

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

For related studies that analyze patient flow in EDs, we refer to Siddharathan et al. (1996), Wang (2004), Huang et al. (2012), and the references therein. Peck and Kim (2010) developed a triage index system, the park index, that accounts for both urgency and patient flow, and described its use in assigning patients to a fast track. Peck et al. (2012) studied possible ways to use the information collected at triage to predict the admission rate to the hospital and discussed how such information can be used for improving bed management, patient flow, and discharge processes. Argon and Ziya (2009) used average waiting time as the performance metric in a general service system with two classes of customers, in which customer classification is imperfect, and showed that prioritizing customers according to the probability of being from the class that should have a higher priority when classification is perfect outperforms any finite-class priority policy. Dobson et al. (2013) developed a heavy-traffic model with an investigator and server interruptions to study physician choice in prioritizing patients.

RIGHTSLINK

Downloaded from informs.org by [129.219.247.33] on 16 August 2014, at 09:39 . For personal use only, all rights reserved.

In the medical literature, Gilboy et al. (2005), FitzGerald et al. (2010), and Ierson and Moskop (2007) provide excellent reviews of the history of the triage process and its development over time. Although triage has been based mainly on urgency, the idea of considering the complexity of patients goes back to World War I mass-casualty triage recommendations: "A single case, even if it urgently requires attention,—if this will absorb a long time,—may have to wait, for in that same time a dozen others, almost equally exigent, but requiring less time, might be cared for. The greatest good of the greatest number must be the rule." (Keen 1917, p. 13). Anticipating the potential of complexity-augmented triage, Vance and Sprivulis (2005) empirically tested the ability of nurses to estimate patient complexity at the time of triage and found that they are able to do this reliably. Vance and Sprivulis (2005) suggested that this type of information could be used to improve patient flow in EDs, although they did not specify how. Finally, it is noteworthy that similar complexity information is also used in mass-casualty triage settings (see, e.g., Jacobson et al. 2012 and the references therein).

3. Modeling the ED

To address the four questions (prioritization, magnitude, sensitivity, and design) posed in §1, we make use of a model of patient flow through the main ED (see Figure 2). We focus our attention on the main ED, which means we do not consider the minority of the patients routed to the resuscitation unit or fast track. A patient's path through the main ED begins with *arrival*, which occurs in a nonstationary stochastic manner. Upon arrival, the patient goes to *triage*, where he or she 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, the patient goes to the *waiting* area until being called by the charge nurse and brought to an examination room. There the patient goes through a stochastic number of *treatment* stages with a physician, which include diagnosis, consultation, and other interactions that are also stochastic in duration. These treatment stages are punctuated by *test* stages during which the patient is unavailable to the physician, which involve testing (MRI, CT scan, x-ray etc.), preparation/processing activities that do not involve the physician, or waiting for test results. The final processing stage after the last physician interaction is *disposition*, in which the patient is either *discharged* to go home or *admitted* to the hospital.

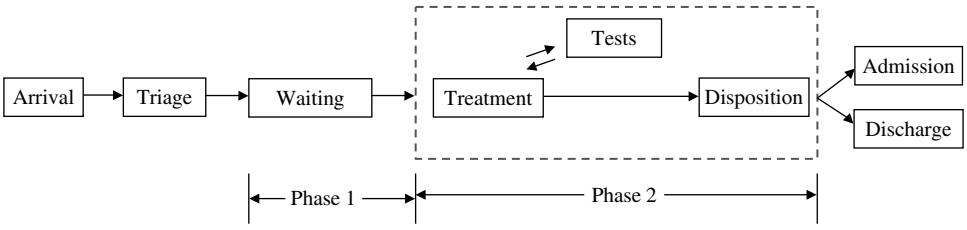
We refer to the time a patient spends after being triaged and before being brought to an examination room as "phase 1," and label the remaining time until disposition as "phase 2." Because they are under observation and care, patients have a lower risk of adverse events during phase 2 than during phase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

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

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

In our framework, a patient's rate of adverse events is influenced by his or her true type $ij \in \mathcal{U} \times \mathcal{C}$, where $i \in \mathcal{U}$

Figure 2 **General Flow of Patients in the Main ED**



RIGHTSLINK

Downloaded from informs.org by [129.219.247.33] on 16 August 2014, at 09:39 . For personal use only, all rights reserved.

is the patient's urgency level, $j \in \mathcal{C}$ is the patient's complexity type, $\mathcal{U} = \{U(\text{Urgent}), N(\text{Nonurgent})\}$ and $\mathcal{C} = \{C(\text{Complex}), S(\text{Simple})\}$. Under urgency-based triage an estimate is made of i, whereas under complexity-augmented triage estimates are made of both i and j.

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

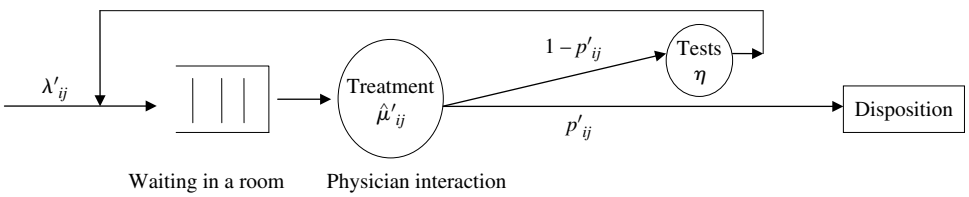
4. Phase 2: Sequencing Patients Within the ED

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

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

For tractability, we assume each interaction with patients classified as ij takes an exponentially distributed amount of time with rate $\hat{\mu}'_{ij}$. We also assume that the physician can preempt an interaction to see a patient of a different class. When a physician returns to a preempted interaction, we assume the physician 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 nonpreemption in phase 2 in §6 using simulation. After each completed interaction, a patient classified as ij may be disposed (discharged home or admitted to the hospital) with probability $p'_{ij} > 0$, or with probability $1 - p'_{ij}$ requires another round of test and treatment. We note that in practice the probability of being disposed may not be constant because it depends on various factors (e.g., progression of pain, the number of past interactions with the physician, revealed test results, etc.). If data on such factors were collected, they could be incorporated into the patient prioritization decision. Since such data do not currently exist, we approximate the number of interactions with the physician by fitting a geometric distribution with constant probability of

Figure 3 Patient Flow After a Patient Is Moved to an Examination Room/Bed (Phase 2 Sequencing)



RIGHTSLINK

Downloaded from informs.org by [129.219.247.33] on 16 August 2014, at 09:39 . For personal use only, all rights reserved.

departure p'_{ij} for patients classified as $ij \in \mathcal{U} \times \mathcal{C}$, which can easily be estimated using the average number of physician-patient interactions for that class. Empirical data from the literature (Graff et al. 1993) suggest that this is a reasonable approximation of reality (see, e.g., Figure 5 in §6).

We model the aggregated test times (including any preparation and waits for results) as i.i.d. exponential random variables. The i.i.d. assumption is justifiable because (a) most testing facilities are shared with units outside the ED and so have workloads not visible to ED personnel, and (b) even if test facility workloads were known, ED personnel could not incorporate them into patient sequencing decisions because they do not know which tests, if any, will be required before examining a patient and/or the patient's prior test results (which begins an interaction). So, for purposes of patient sequencing at least, test time delays look like i.i.d. random variables to the physician. Modeling such delays as exponential is reasonable because (1) the waiting time distribution in many queueing systems is exponential or nearly so, and (2) at least in UMHSED, physicians do not make use of the "age" information that a delay distribution with nonconstant failure rate would provide. (The assumptions above give us a tractable $\cdot/M/\infty$ approximation of test time delays from the perspective of an ED physician. Although this approximation is useful to gain insights into phase 2 prioritization, it does not facilitate examination of the potential for coordinating ED decisions with the real-time status of test facilities. Studies of test facilities upon which a more detailed model with which to consider this possibility could be constructed include Green et al. 2006a, Patrick et al. 2008, and Batt and Terwiesch 2012.)

To keep our analytical model tractable, the aggregate test delay times are further assumed to be a generic "test" with mean time η^{-1} that is the same across different patient classes. However, we relax this assumption and allow patient class specific test delays in our simulation model of §6. Finally, we note that in most EDs physicians do not update patients' triage classes for various reasons including those related to liability. Hence, consistent with practice, we assume patient classifications are made at triage and are not updated during the phase 2 service process. We refer to the representation of phase 2 of the ED service with above assumptions as the *simplified phase 2 model with dynamic arrivals*.

Because each physician is dedicated to his or her own slate of patients, we focus on a single physician's decision of who to see next. To this end, we let $\underline{x} = (x_{ij})_{ij\in\mathcal{U}\times\mathcal{C}}$ (respectively, $\underline{y} = (y_{ij})_{ij\in\mathcal{U}\times\mathcal{C}}$) represent the error-impacted number of patients of each class available (not available) for the physician visit. With these, we can define the state of the system at any point in time, t, by the vector $(\underline{x}(t), \underline{y}(t)) \in \mathbb{Z}_+^4 \times \mathbb{Z}_+^4$, and model the process $\{(\underline{x}(t), \underline{y}(t)): t \geq 0\}$ as a continuous time Markov chain (CTMC). Because here we are considering an exogenous arrival process that represents underload conditions (we will consider overload conditions later), the physician's capacity and the number of beds are not binding, and hence we do not impose any explicit bounds on $(\underline{x}(t), \underline{y}(t))$. As an underloaded system, we may assume the parameters of the system are such that the underlying CTMC is stablizable; that is, there exists at least one policy under which the risk of adverse events is finite. However, because of the Poisson arrivals and the pure delay model for test times, the transition rates are not bounded, and hence we cannot use the uniformization method of Lippman (1975) to formulate a discrete time equivalent of the CTMC in which the times between consecutive events are i.i.d. (for all states). So instead, we construct a sequence of controlled CTMC's (CCTMC's) with an increasing but bounded sequence of (maximum) transition rates converging to the original CCTMC. We do this by replacing the $\cdot/M/\infty$ model of test process 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 does not need to be captured in the system's state. Another novel aspect of our approach is that we truncate the transition rates instead of truncating the state space, thereby avoiding the artificial boundary effects that usually distort the optimal policy. Since the transition rates in the CTMC indexed by k (for all k) are bounded by $\psi_k = \max_{ij\in\mathcal{U}\times\mathcal{C}} \hat{\mu}'_{ij} + 4k\eta + \sum_{ij\in\mathcal{U}\times\mathcal{C}} \lambda'_{ij} < \infty$, 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.

For the system indexed by k, the optimal rate of adverse events under a patient classification based on both sets $\mathcal{U}$ and $\mathcal{C}$, $R^{k*} = \inf_{\pi\in\mathbf{\Pi}} R_\pi^{\mathcal{U}\cup\mathcal{C}}$ (where $\mathbf{\Pi}$ denotes the set of all admissible Markovian policies), and the optimal physician behavior can be derived from the following average cost optimality equation:

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

RIGHTSLINK

Downloaded from informs.org by [129.219.247.33] on 16 August 2014, at 09:39 . For personal use only, all rights reserved.

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 0 elsewhere, a is an action determining which patient class to serve, and $\mathscr{A}(\underline{x}) = \{ij \in \mathscr{U} \times \mathscr{C}: x_{ij} > 0\} \cup \{0\}$ is the set of feasible actions (class 0 represents the idling action) when the error-impacted number of patients of each class in the examination rooms is $\underline{x}$.

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

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

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

As a further check on the robustness of this prioritization result, we consider an alternate model in which arrivals to phase 2 are not dynamic. Specifically, we assume that all patients for a given time interval (e.g., the afternoon rush period) arrive to the ED waiting room at once and the objective is to clear them out as quickly as possible to minimize LOS and ROAE. We further assume that ED physicians can treat any patient in the system (i.e., there are no constraints on the number of beds or the number of patients per physician). However, we note that because interactions with patients are time consuming and because physicians revisit patients already in phase 2, this model will limit the rate of patient flow into phase 2 even without these constraints. We label this the simplified phase 2 model with static arrivals. In contrast with the simplified phase 2 model with dynamic arrivals, which is representative of the ED under light load conditions, this model is representative of the ED under heavy load conditions that create a backlog of patients. In Online Appendix D, we show that the phase 2 sequencing policy of Theorem 1 remains optimal under these very different modeling conditions. The suggestion is that the cost/time balance struck by the $c\mu$-type rule of Theorem 1 is robustly effective in phase 2 for various arrival processes, even ones that may be potentially endogenous to the priority rule. Our simulation studies also support this observation.

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

5. Phase 1: Sequencing Patients Into the ED

To create a simplified model that captures the essential dynamics of sequencing patients into the ED, we represent the dashed area in Figure 2 (i.e., phase 2) as a single-stage aggregated service node with a single "super server" that represents the aggregate ED capacity. Because our phase 2 analysis indicated that simple patients with a given urgency level should be prioritized over complex ones within the ED and, by definition, complex patients have on average more interactions with physicians, it follows that simple patients should have a higher aggregate/effective service rate than complex ones. Specifically, we suppose patients of type $ij \in \mathscr{U} \times \mathscr{C}$ have i.i.d. service times (i.e., the total time spent in phase 2) that follow a general distribution, $F_{ij}(s)$ with first moment $1/\mu_{ij}$, where $\mu_{iC} \leq \mu_{iS}$ for all $i \in \mathscr{U}$, and a finite second moment. (Note that that these service rates can be computed using sojourn times from a simulation or Markov chain

RIGHTSLINK

Downloaded from informs.org by [129.219.247.33] on 16 August 2014, at 09:39 . For personal use only, all rights reserved.

analysis under the optimal phase 2 service policy.) For tractability, we model the arrival of patients of type $ij \in \mathcal{U} \times \mathcal{C}$ to the ED as a Poisson process with rate λ_{ij}. As we did in our simplified phase 2 models, we assume patients of type ij are subject to adverse events, which occur according to a Poisson process with intensity θ_{ij}, where $\theta_{Uj} \geq \theta_{Nj}$ for all $j \in \mathcal{C}$. These assumptions lead to a tractable model of phase 1; however, many of them will be relaxed later (see, e.g., Online Appendix C and §6).

5.1. Urgency-Based Triage

We first consider current practice in most EDs in which patients in the main ED (levels 2 and 3) are classified solely based on urgency, and use our simplified phase 1 model to examine decisions for sequencing patients into the ED. We start with the case of perfect classification and then consider the case of stochastic misclassification. When patients can be perfectly classified as either urgent (U) or nonurgent (N), the arrival rates for Us and Ns are $\lambda_U = \sum_{j\in\mathcal{C}} \lambda_{Uj}$ and $\lambda_N = \sum_{j\in\mathcal{C}} \lambda_{Nj}$, respectively. Similarly, the average service times for Us and Ns are $1/\mu_U = \sum_{j\in\mathcal{C}} (\lambda_{Uj}/\lambda_U)(1/\mu_{Uj})$ and $1/\mu_N = \sum_{j\in\mathcal{C}} (\lambda_{Nj}/\lambda_N)(1/\mu_{Nj})$, respectively. Furthermore, from known results for nonpreemptive priority queues (see, e.g., Cobham 1954) the average waiting (queue) time of the kth priority class is

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

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

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

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

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

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

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

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

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

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

where $\rho'_N = \lambda'_N/\mu'_N$. We summarize the implications of these results in the following proposition. Note that part (i) of this proposition coincides with the "expected $c\mu$" and highest signal first (HSF) policy discussed in Argon and Ziya (2009). However, because of differences between our setting and theirs (most importantly, Argon and Ziya (2009) assume a continuous signal from customers that indicates the probability of misclassification, whereas our approach uses average misclassification error rates that can be estimated from data), we provide an independent proof based on the above results.

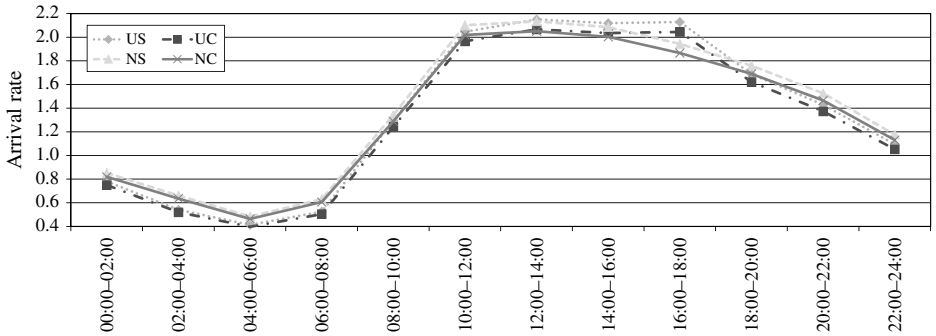
Proposition 1 (Phase 1 Prioritization—Urgency-Based Triage). *In the simplified phase* 1 *model with imperfect urgency-based classification*, (i) *The best* (*static*) *priority rule is to prioritize U patients if* $\theta'_U\mu'_U \geq \theta'_N\mu'_N$; *otherwise, prioritize N patients*. (ii) *The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

Part (i) of Proposition 1 shows how triage misclassification errors can be incorporated into the optimal priority rule. Part (ii) of Proposition 1 shows that if the misclassification rate is high enough, the optimal priority rule prioritizes N patients. However, empirical studies have observed misclassification levels γ_N and γ_U to be in the range 9%–15% depending on the level of triage nurse experience (Hay et al. 2001). Thus, if,

RIGHTSLINK

Downloaded from informs.org by [129.219.247.33] on 16 August 2014, at 09:39 . For personal use only, all rights reserved.

Figure 4 Class Dependent Arrival Rates to the ED for an Average Day (Obtained from a Year of Data in UMHSED)



as we expect, prioritizing urgent patients is optimal when there is no misclassification error, prioritizing them remains optimal even under practical levels of misclassification errors. This confirms that prioritizing level 2 patients over level 3 patients, as is typically done in the main ED, is reasonable. However, we note that there is wide variance of complexity among level 2 and level 3 patients (see, e.g., Vance and Sprivulis 2005 and Figure 4). Simply prioritizing level 2 patients over level 3 patients may be significantly suboptimal relative to a policy that considers complexity. We investigate this issue in the next section.

5.2. Complexity-Augmented Triage

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

To evaluate the performance of complexity-augmented triage when classification is imperfect, we again let γ_U and γ_N denote the misclassification error rates with respect to set $\mathcal{U}$. Similarly, we let γ_C and γ_S denote the misclassification error rates with respect to set $\mathcal{C}$, γ_C denote the probability that a C patient is classified as an S, and γ_S denote the probability that an S patient is classified as a C. We assume the misclassification probabilities with respect to $\mathcal{U}$ and $\mathcal{C}$ are independent because (a) the data show that the assessments themselves are uncorrelated, indicating that urgency and complexity are medically separable questions (e.g., Figure 4 indicates that an urgent patient is almost equal likely to be simple or complex (and vice versa)), and (b) multiple nurses perform triage, thereby limiting the extent of any systematic biases in misclassifications.

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

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

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

Similarly, if $\underline{\theta}'$ and $\underline{\mu}'$ denote the vector of error-impacted adverse event and service rates, we have $\underline{\theta}'^T = (A(\underline{\lambda} \times \underline{\theta})^T)/\underline{\lambda}'$ and $(\underline{1}/\underline{\mu}')^T = (A(\underline{\lambda}/\underline{\mu})^T)/\underline{\lambda}'$, where $\underline{1} = (1, 1, 1, 1)$ and operators "$\times$" and "$/$" are component-wise multiplication and division, respectively. With these, the waiting times for each customer class under an imperfect $\mathcal{U} \cup \mathcal{C}$ classification can be computed using (2) with rates replaced with their transformed error impacted counterparts. This model permits us to show the following.

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

RIGHTSLINK

Downloaded from informs.org by [129.219.247.33] on 16 August 2014, at 09:39 . For personal use only, all rights reserved.

information into phase 1 sequencing. However, the indices used are linearly transformed to incorporate misclassifications. Thus, this result can be viewed as an extension of the $c\mu$ rule under imperfect information. Furthermore, Proposition 2(i) shows precisely when the optimal priority rule will be different from the optimal rule without misclassification. For instance, if misclassification rates are high enough, it can be better to prioritize nonurgent patients over urgent ones. However, at the error levels observed in the previously cited studies, the implication of Proposition 2(i) is to prioritize patients in the order: US, UC, NS, NC, which coincides with the priority rule we found to be optimal in phase 2. Proposition 2(ii) begins to address the magnitude question raised in the introduction by suggesting that complexity-augmented triage outperforms urgency-based triage, provided that the optimal priority rule is implemented. Although this result may seem intuitive because of the additional information collected at triage, we note that the additional information is subject to errors, so this conclusion is not obvious. Nevertheless, Proposition 2(ii) shows that, implemented correctly, imperfect complexity-augmented information *always* improves the ED performance regardless of the misclassification levels. Moreover, it should be noted that information collection involves only simple estimations of whether two or more interactions are needed with a physician, which adds minimal time to the triage process. Whereas priority rules are greedy and usually suboptimal, Proposition 2(iii) confirms that they are optimal in this setting. The surprise is that it is never optimal to idle in anticipation of a high priority patient when only low priority patients are available, even though the model disallows preemption. Similar results for the $c\mu$ rule but without misclassifications are presented in Kakalik and Little (1971). Finally, part (iii) of Proposition 2 states that a dynamic (i.e., state-dependent) priority policy cannot beat the greedy and simple state-independent policy presented in part (i).

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

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

This implies that complexity-augmented triage is most beneficial in EDs with (i) high utilization, (ii) high heterogeneity in the average service time of simple and complex patients, (iii) equal fractions of simple and complex patients, or (iv) low complexity classification error rates.

5.3. Patient Flow Design Using Complexity and Urgency Information

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

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

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

The intuition behind the above result is that matching capacity to workload in the different streams diminishes the effect of different service times among simple and complex patients. Hence, the difference in the

RIGHTSLINK

Downloaded from informs.org by [129.219.247.33] on 16 August 2014, at 09:39 . For personal use only, all rights reserved.

intensity of adverse events between urgent and nonurgent patients becomes the dominant factor in selecting the best streaming design. Because complexity streaming prioritizes patients in both streams according to urgency, it is more effective than urgency streaming. Although we have derived this insight using a single server model, we note that the single server assumption is not essential to the comparison. First, it can be easily seen from the well-known Sakasegawa equation (see, e.g., Hopp and Spearman 2008, pp. 290–291) that with the same utilization, the queueing time of a G/G/k queueing system and that of an "equivalent" G/G/1 become equal as the system's utilization approaches 1. Since utilization in the ED is typically high, we expect the single server assumption to provide a good approximation. Second, our numerical comparisons show that even with utilization as low as 60%, the conclusion that complexity-based streaming is superior to urgency-based streaming is not altered by the number of servers. Finally, in §6.4, we use hospital data and a simulation model with multiple physicians, beds, etc. to further confirm the superiority of complexity streaming.

6. Simulation Analysis of Complexity-Augmented Triage

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

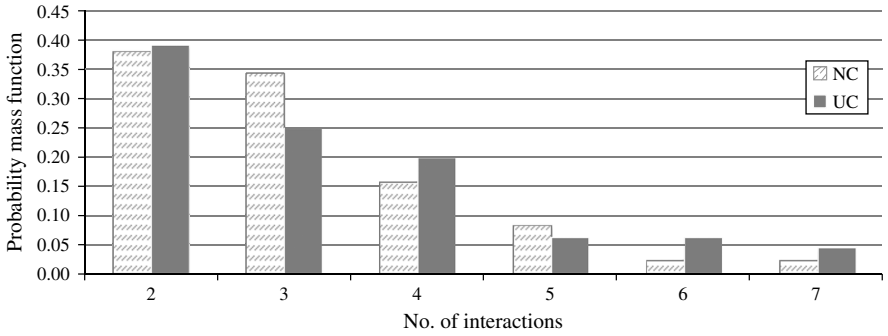
Patient Classes. At the time of triage, patients are classified according to both urgency (urgent or nonurgent) and complexity (simple or complex). For modeling purposes, we omit the resuscitation unit (RU) and fast track (FT) classifications, shown in Figure 1(b), since these patients are typically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

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

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

RIGHTSLINK

Figure 5 Probability Mass Function of the Number of Class-Based Interactions for Complex Patients



Downloaded from informs.org by [129.219.247.33] on 16 August 2014, at 09:39 . For personal use only, all rights reserved.

classes. The simulated service process is considered to be noncollaborative (since an ED physician rarely transfers his or her patients to another physician) and also nonpreemptive.

Physician-Patient Assignments and Priorities. Patients are brought back from the waiting area to exam rooms whenever a room becomes available according to a phase 1 sequencing rule. When a physician becomes available, and has fewer than his or her maximum number of patients (seven is typical), the physician chooses the next patient from those available based on a phase 2 sequencing rule, which can make use of information generated at triage. We assume that when urgency-based triage is used, U patients get priority over N patients in both phases 1 and 2. If complexity-augmented triage is used, patients are prioritized in both phases according to the priority ordering US, UC, NS, NC (ranked from high to low priority), which we found to be optimal in the simplified ED models discussed previously. When a patient is brought back to an examination room, we assume that the patient is assigned to the physician with the lowest number of patients. If all physicians are handling their limit of seven 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 four physicians in our base case scenario, which are representative of a medium sized ED. But we perform sensitivity analysis to examine the effect of number of both beds and physicians on the benefit of complexity-augmented triage. We also consider cases with nonstationary staffing in order to examine the effect of better matching staffing to the demand profile. We consider test facilities (ancillary services) as exogenous resources (i.e., test times are independent of the volume of ED patients) because these facilities typically handle many other patients besides those from the ED. Hence, tests and waiting for their reports result in various exogenous "delays," which were approximated with class dependent exponential random variables and were estimated to have mean "delay" times roughly 2.5 times longer for complex patients than for simple patients based on UMHSED data.

Adverse Events. Adverse events are simulated using point processes with stationary rates that depend on patient class and phase of service. Specifically, in our base case we use class and phase dependent Poisson processes and assume that (i) U patients have a higher intensity of adverse events than N patients, and (ii) the intensity of adverse events decreases by 60% when patients move from the waiting room (phase 1) to an examination room (phase 2). (The 60% number is a physician estimate based on the impact of more careful monitoring and care within the ED; however, we give sensitivity analysis in Online Appendix E that shows the main conclusions are robust to this estimate.) As in our previous models, we do not consider fatal events that would terminate the adverse events counting process, since the impact of these rare events on our objective function is extremely small.

Runs. The simulation was written in C++ and made use of a cyclo-stationary model (see, e.g., Gardner et al. 2006 for a complete review of cyclo-stationarity) with a period of a week. Each data point was obtained for 5,000 replications of one week, where each replication was preceded by a warm-up period of one week. This was observed to be sufficient because correlations in the ED flow are very small for spans of two or more days because of the fact that EDs generally clear out overnight. The number of replications (5,000) was chosen to achieve reliable confidence intervals that are tight enough to be omitted from our data presentations.

6.1. Performance of Complexity-Augmented Triage

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

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

RIGHTSLINK

Downloaded from informs.org by [129.219.247.33] on 16 August 2014, at 09:39 . For personal use only, all rights reserved.

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

The smaller effect of phase 2 sequencing compared to that of phase 1 prioritization is mainly due to the fact that, under the conditions of our base case, physicians in phase 2 often do not have many available patients from which to choose. This is because (a) patients are unavailable for considerable amounts of time while being tested and waiting for test results, and (b) each physician handles only a limited number of patients simultaneously (with an upper bound of seven). However, in EDs 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, three physicians, and a maximum number of 10 patients per physician. Under these conditions, if phase 2 sequencing is done according to SIRO for both the urgency-based and complexity-augmented triage systems, then complexity-augmented triage achieves improvements of 8.6% and 6.2% in ROAE and LOS, respectively, relative to urgency-based triage. In contrast, if the urgency-based triage system prioritizes patients in phase 2 by urgency ($U > N$) and the complexity-augmented triage system prioritizes patients in phase 2 by complexity and urgency ($US > UC > NS > NC$), then complexity-augmented triage achieves improvements of 13.1% and 9.10% in ROAE and LOS, respectively, relative to urgency-based triage. This leads us to the following:

Observation 2. In EDs 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. The Effect of ED Resource Levels

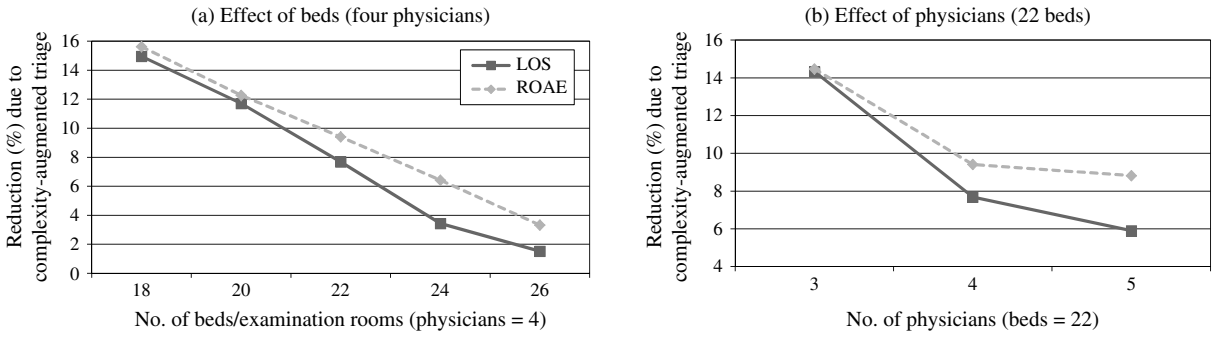
Proposition 3 predicts that increased utilization of resources (i.e., either physicians or examination rooms) should favor complexity-augmented triage. Figure 6 illustrates the percentage improvement in terms of ROAE and LOS from using complexity-augmented triage over urgency-based triage for varying numbers of examination rooms and physicians. From this figure we observe the following:

Observation 3. The benefit of complexity-augmented triage is greater in EDs with higher bed and/or physician utilization.

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

Nonstationary Staffing. Because EDs typically adjust staffing to follow workload, at least to some extent, we now consider two cases of nonstationary staffing: (i) reducing the staffing level during off-peak hours, and (ii) reducing the staffing level during off-peak hours while increasing staffing during peak hours (redistributing the current workforce). To examine these cases, we consider two alternate scenarios that modify our base case assumption of four physicians at all times: (i) four physicians during peak demand times (12-hour shifts) and three physicians otherwise, and (ii) six

Figure 6 The Effect of Resources (Beds and Physicians) on the Benefit of Complexity-Augmented Triage over the Current Practice of Urgency-Based Triage



RIGHTSLINK

Downloaded from informs.org by [129.219.247.33] on 16 August 2014, at 09:39 . For personal use only, all rights reserved.

physicians during peak times (12-hour shifts) and two otherwise, so there is no net change in labor hours. Under scenario (i), the complexity-augmented triage achieves 11.5% and 11.0% improvements in ROAE and LOS, respectively, compared to urgency-based triage. Under scenario (ii), these numbers are 8.8% and 6.9%. Hence, the improvements relative to the 9.4% and 7.4% improvements of the base case shown in Figure 6(b) are larger under scenario (i) but not under scenario (ii). The reason is that scenario (i) increases utilization during off peak hours, which we have already shown enhances the benefits of complexity-augmented triage. But scenario 2 increases utilization during peak hours while decreasing it during off peak hours. Since overall performance is dominated by the peak hours, during which most congestion occurs, this results in a net decrease in the benefits of complexity-augmented triage. Nevertheless, since it is not economical to entirely eliminate high utilization periods in the ED, the benefits of complexity-augmented triage will be reduced but not eliminated with staffing that better matches the demand profile.

Bed-Block Phenomenon. We can use the results of Figure 6 to predict the effect of the ED bed-block phenomenon, in which ED patients admitted to the hospital cannot be transferred to their inpatient unit because of unavailability of beds. By tying up beds in the ED to board admitted patients, bed-block reduces the effective number of ED beds, and hence, increases their utilization. Therefore, from Figure 6 and Observation 3, we can expect complexity-augmented triage to yield greater benefits in EDs with higher bed-block/boarding times. For more detailed discussion of the effect of ED bed-block on patient flow design, we refer interested readers to Saghafian et al. (2012) and the related references therein.

6.3. The Effect of Misclassification

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

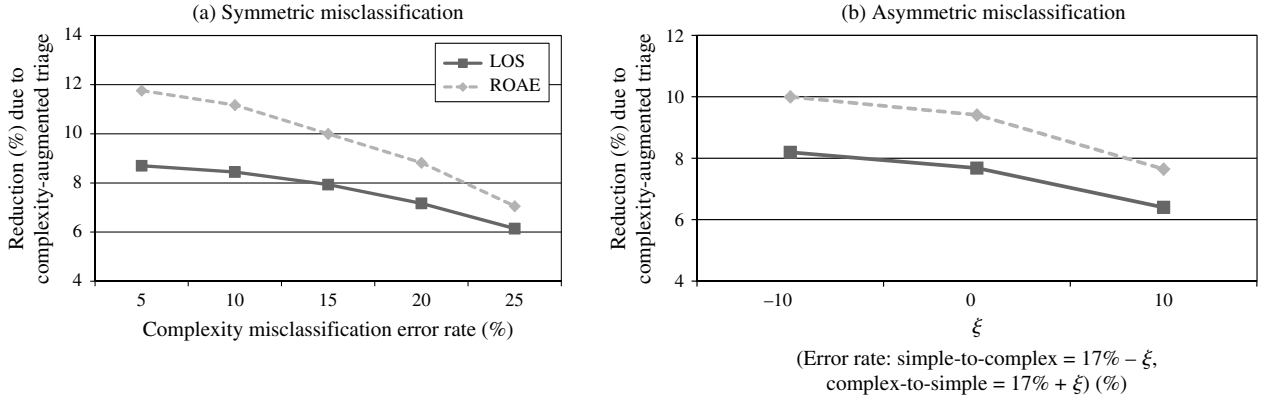
Observation 4. The benefit of complexity-augmented triage is relatively robust to complexity misclassification errors. However, complex-to-simple misclassifications are slightly more harmful than simple-to-complex misclassifications.

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

6.4. Complexity Streaming Patient Flow Design

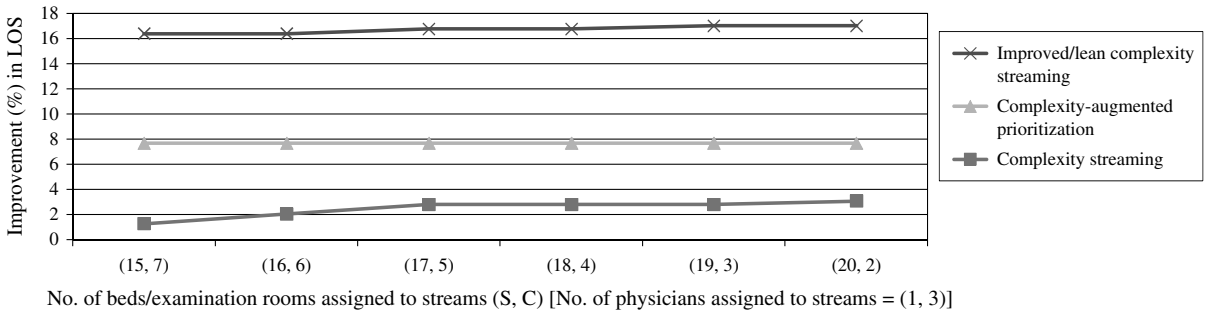
Finally, we return to the question of whether patient complexity information is most valuable in prioritizing or streaming patients. To do this, we examine a *complexity streaming* design in which patients are divided into two streams: one for patients triaged as simple (S) and one for those triaged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

Figure 7 **The Effect of Complexity Misclassification Error Rates on the Benefit of a Complexity-Augmented Triage (Compared to an Urgency-Based Only Triage)**



RIGHTSLINK

Figure 8 Performance of Different Patient Flow Designs Compared to the Current Practice (Urgency Prioritization)



of type S may enter a C bed if no other C patients are waiting, and vice-versa. This type of flexible allocation mechanism is referred to as "virtual streaming" in Saghafian et al. (2012), which demonstrates it to be important in disposition-based streaming protocols. Finally, we assume that patients in both streams and in both phases 1 and 2 are prioritized according to their urgency level.

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

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

Observation 5. It is better to use complexity information for streaming and urgency information for prioritizing than using urgency information for streaming and complexity information for prioritizing (5.7% and 4.8% improvements in ROAE and LOS, respectively).

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

7. Conclusion

In this paper, we propose a new triage system for ED practice in which patients are classified on the basis of complexity, as well as urgency. Our results suggest that, compared to current urgency-based triage systems, complexity-augmented triage can significantly improve ED performance in terms of both patient safety (ROAE) and operational efficiency (LOS), even if patient classification is subject to error. We find that a simple and fast classification scheme, which defines patients to be simple if they require only a single interaction (and complex otherwise) works very well as the basis for complexity-augmented triage because it results in (1) a nearly even split between simple and complex patients and (2) a substantial difference between average treatment time of complex and simple patients.

Our analyses indicate that complexity-augmented triage can yield substantial safety and efficiency improvements even if complexity information is only used to prioritize patients up to entry into the examination rooms (phase 1). Furthermore, in EDs where physicians have a significant amount of choice about

Downloaded from informs.org by [129.219.247.33] on 16 August 2014, at 09:39 . For personal use only, all rights reserved.

RIGHTSLINK

Downloaded from informs.org by [129.219.247.33] on 16 August 2014, at 09:39 . For personal use only, all rights reserved.

what patient to see next within examination rooms (phase 2), we find that complexity information gathered at triage can yield additional benefits by facilitating internal sequencing decisions. For both phase 1 and phase 2, the benefit of complexity-augmented triage is greatest in EDs with high physician and/or examination room utilization. Since EDs are widely overcrowded, our results suggest that complexity-augmented triage is an effective way for EDs to improve safety and reduce congestion without adding expensive human or physical capacity.

We also investigate a new patient flow design, in which complexity-augmented triage information is used to separate simple and complex patients into two streams. Our results suggest that it is more effective to stream patients based on their complexity and then prioritize them within each based on their urgency than it is to stream them according to urgency and prioritize them according to complexity. Streaming based on complexity also facilitates implementation of lean methods in the "simple" patient stream, which can take advantage of complexity-augmented triage information to achieve even greater gains. If these gains are substantial enough, such complexity streaming can yield significant additional benefits.

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

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/msom.2014.0487.

Acknowledgments

Partial support was provided by the National Science Foundation [Grants CMMI-1068638 and CMMI-1233095], and the National Institutes of Health Clinical and Translational Science Award (CTSA) [Grant UL1TR000433]. The funding organizations had no role in the design or conduct of this research.

References

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

Baker M, Clancy M (2006) Can mortality rates for patients who die within the emergency department, within 30 days of discharge from the emergency department, or within 30 days of admission from the emergency department be easily measured? *Emergency Medical J.* 23(8):601–603.

Batt RJ, Terwiesch C (2012) Doctors under load: An empirical study of state-dependent service times in emergency care. Working paper, The Wharton School, University of Pennsylvania, Philadelphia.

Ben-Tovim DI, Bassham JE, Bennett DM, Dougherty ML, Martin MA, O'Neill JL, Sincock JL, Szwarcboard MG (2008) Redesigning care at the Flinders Medical Centre: Clinical process redesign using "lean thinking." *Medical J. Australia* 188(6): 27–31.

Brennan TA, Leape LL, Laird NM, Hebert L, Localio AR, Lawthers AG, Newhouse JP, Weiler PC, Hiatt HH (1991) Incidence of adverse events and negligence in hospitalized patients. *New England J. Medicine* 324(6):370–376.

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

Clark JR, Huckman RS (2012) Broadening focus: Spillovers, complementarities, and specialization in the hospital industry. *Management Sci.* 58(4):708–722.

Cobham A (1954) Priority assignment in waiting line problems. *J. Oper. Res. Soc. Amer.* 2:70–76.

Cobham A (1955) Priority assignment—A correction. *J. Oper. Res. Soc. Amer.* 3:547.

Cox DR, Smith WL (1961) *Queues* (Methuen & Co., London).

Diercks DB, Roe MT, Chen AY, Peacock WF, Kirk JD, Pollack CV Jr, Gibler WB, Smith SC Jr, Ohman M, Peterson ED (2007) Prolonged emergency department stays of non-ST-segment-elevation myocardial infarction patients are associated with worse adherence to the American College of Cardiology/American Heart Association guidelines for management and increased adverse events. *Ann. Emergency Medicine* 50(5):489–496.

Dobson G, Tezcan T, Tilson V (2013) Optimal workflow desicions for inestigators in systems with interruptions. *Management Sci.* 59(5):1125–1141.

Fernandes CM, Tanabe P, Gilboy N, Johnson LA, McNair RS, Rosenau AM, Sawchuk P, et al. (2005) Five-level triage: A report from the ACEP/ ENA five-level task force. *J. Emergency Nursing* 31(1):39–50.

FitzGerald G, Jelinek GA, Scott D, Gerdtz MF (2010) Emergency department triage revisited. *Emergency Medicine J.* 27(2): 86–92.

Gardner WA, Napolitanob A, Paura L (2006) Cyclostationarity: Half a century of research. *Signal Processing* 86(4):639–697.

Gilboy N, Tanabe P, Travers DA, Rosenau AM, Eitel DR (2005) *Emergency Severity Index, Version 4: Implementation Handbook* (Agency for Healthcare Research and Quality, Rockville, MD).

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

Green L, Savin S, Wang B (2006a) Managing patient service in a diagnostic medical facility. *Oper. Res.* 54(1):11–25.

Green LV, Soares J, Giglio JF, Green RA (2006b) Using queuing theory to increase the effectiveness of emergency department provider staffing. *Acad. Emergency Medicine* 13(1):61–68.

Hay E, Bekerman L, Rosenberg G, Peled R (2001) Quality assurance of nurse triage: Consistency of results over three years. *Amer. J. Emergency Medicine* 19(2):113–117.

Hopp WJ, Spearman ML (2008) *Factory Physics*, 3rd ed. (McGraw-Hill, Burr Ridge, IL).

Huang J, Carmeli B, Mandelbaum A (2012) Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. Working paper, National University of Singapore, Singapore.

RIGHTSLINK

Ierson KV, Moskop JC (2007) Triage in medicine, part I: Concept, history, and types. *Ann. Emergency Medicine* 49(3):275–281.

Jacobson EU, Argon NT, Ziya S (2012) Priority assignment in emergency response. *Oper. Res.* 60(4):813–832.

Kakalik JS, Little JDC (1971) Optimal service policy for the $M/G/1$ queue with multiple classes of arrival. Report, RAND Corporation, Santa Monica, CA.

KC DS, Terwiesch C (2011) The effects of focus on performance: Evidence from California hospitals. *Management Sci.* 57(11): 1897–1912.

Keen WW (1917) *The Treatment of War Wounds* (W.B. Saunders, Philadelphia).

Kronick SL, Desmond JS (2009) Blink: Accuracy of physician estimates of patient disposition at the time of ED triage. *SAEM Midwest Regional Meeting, Ann Arbor, MI.*

Lewis PAW, Shedler GS (1979) Simulation of nonhomogenous Poisson processes by thinning. *Naval Res. Log. Quart.* 26(3):403–413.

Lippman S (1975) Applying a new device in the optimization of exponential queueing system. *Oper. Res.* 23(4):687–710.

Liu SW, Thomas SH, Gordon JA, Weissman J (2005) Frequency of adverse events and errors among patients boarding in the emergency department. *Acad. Emergency Medicine* 12:49–50.

Moskop JC, Ierson KV (2007) Triage in medicine, part II: Underlying values and principles. *Ann. Emergency Medicine* 49(3): 282–287.

Patrick J, Putterman ML, Queyranne M (2008) Dynamic multipriority patient scheduling for a diagnostic resource. *Oper. Res.* 56(6): 1507–1525.

Peck JS, Benneyan JC, Nightingale DJ, Gaehde SA (2012) Predicting emergency department inpatient admissions to improve same-day patient flow. *Acad. Emergency Medicine* 19(9):1045–1052.

Peck JS, Kim S-G (2010) Improving patient flow through axiomatic design of hospital emergency departments. *CIRP J. Manufacturing Sci. Tech.* 2(4):255–260.

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

Saghafian S, Van Oyen MP, Kolfal B (2011) The "W" network and the dynamic control of unreliable flexible servers. *IIE Trans.* 43(12):893–907.

Siddharathan K, Jones WJ, Johnson JA (1996) A priority queueing model to reduce waiting times in emergency care. *Internat. J. Health Care Quality Assurance* 9(5):10–16.

Tcha D-W, Pliska SR (1977) Optimal control of single-server queuing networks and multi-class $M/G/1$ queues with feedback. *Oper. Res.* 27(2):248–258.

van der Zee SP, Theil H (1961) Priority assignment in waiting-line problems under conditions of misclassification. *Oper. Res.* 9(6):875–885.

Van Mieghem JA (1995) Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *Ann. Appl. Prob.* 5(3):809–833.

Vance J, Sprivulis P (2005) Triage nurses validly and reliably estimate emergency department patient complexity. *Emergency Medicine Australasia* 17:382–386.

Wang Q (2004) Modeling and analysis of high risk patient queues. *Eur. J. Oper. Res.* 155(2):502–515.

Downloaded from informs.org by [129.219.247.33] on 16 August 2014, at 09:39 . For personal use only, all rights reserved.

RIGHTSLINK