Management Science

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

Patient Triage and Prioritization Under Austere Conditions

Zhankun Sun, Nilay Tanık Argon, Serhan Ziya

To cite this article:
Zhankun Sun, Nilay Tanık Argon, Serhan Ziya (2017) Patient Triage and Prioritization Under Austere Conditions. Management Science

Published online in Articles in Advance 16 Oct 2017

. https://doi.org/10.1287/mnsc.2017.2855

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–19
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Patient Triage and Prioritization Under Austere Conditions

Zhankun Sun,[a] Nilay Tanık Argon,[b] Serhan Ziya[b]

[a] Department of Management Sciences, College of Business, City University of Hong Kong, Kowloon, Hong Kong; [b] Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, North Carolina 27599

Contact: zhankun.sun@cityu.edu.hk, http://orcid.org/0000-0002-1576-3372 (ZS); nilay@unc.edu, http://orcid.org/0000-0002-6814-0849 (NTA); ziya@unc.edu, http://orcid.org/0000-0003-1558-6051 (SZ)

Received: January 7, 2014
Revised: June 2, 2015; August 19, 2016; May 2, 2017
Accepted: May 13, 2017
Published Online in Articles in Advance: October 16, 2017

https://doi.org/10.1287/mnsc.2017.2855

Abstract. In war zones and economically deprived regions, because of extreme resource restrictions, a single provider may be the sole person in charge of providing emergency care to a group of patients. An important question the provider faces under such circumstances is whether or not to perform triage and how to prioritize the patients. By choosing to triage a particular patient, the provider can determine the health condition and thus the urgency of the patient, but that will come at the expense of delaying the actual service (stabilization or initial treatment) for that patient as well as all the other patients. Motivated by this problem, which also arises in other service contexts, we consider a service system where finitely many patients, all available at time zero, belong to one of the two possible triage classes, where each class is characterized by its waiting cost and expected service time. Patients' class identities are initially unknown, but the service provider has the option to spend time on triage to determine the class of a patient. Our objective is to identify policies that balance the time spent on triage with the time spent on service by minimizing the total expected cost. We provide a complete characterization of the optimal dynamic policy and show that the optimal dynamic policy that specifies when to perform triage is determined by a switching curve, and we provide a mathematical expression for this curve. One insight that comes out of this characterization is that the server should start with performing triage when there are sufficiently many patients and never perform triage when there are few patients. Finally, we carry out a numerical study in which we demonstrate how one can use our mathematical results to develop policies that can be used in mass-casualty triage and prioritization, and we find that there are substantial benefits to using one of these policies instead of the simpler benchmarks.

History: Accepted by Assaf Zeevi, stochastic models and simulation.
Funding: This work is supported by the National Science Foundation [Grants CMMI0927607, CMMI1234212, and CMMI1635574].
Supplemental Material: The online supplement is available at https://doi.org/10.1287/mnsc.2017.2855.

Keywords: triage • priority scheduling • clearing system • Markov decision processes

1. Introduction

Patient triage and prioritization decisions in daily emergencies as well as mass-casualty events primarily aim to make the best use of limited medical resources in an effort to save the lives of as many people as possible and more broadly mitigate the events' negative impact on patients' health. These decisions are highly important regardless of whether they are made in response to daily emergencies or a mass-casualty event, the number of patients seeking treatment, the size of the event, or more generally how limited resources are. However, typically, different factors are at play and different considerations arise depending on the degree to which resources are limited. This paper is concerned with patient triage and prioritization decisions under extremely resource-restricted conditions. Specifically, we focus on settings in which demand for skilled medical providers far surpasses the available supply in close vicinity. In most cases, such conditions are temporary and caused by incidents such as an armed attack, bombing, or an accident, but they might also be chronic as a result of economic deprivation in a region.

A typical emergency response effort to a mass-casualty event in an urban area may involve a team of medical personnel having a range of capabilities and responsibilities ranging from patient triage to resuscitation, transportation, and on-site treatment and surgery. Under such conditions, because different individuals have different skill sets and the number of providers on the scene is relatively large, patient triage and patient treatment can be done by separate groups of individuals in parallel. However, in the case of incidents that occur at geographically isolated locations, battlefields, or military missions that result in multiple life-threatening injuries, a single physician, nurse, or paramedic might find himself/herself as the only person having the skills to deliver proper treatment—at least temporarily—to the injured (Mabry and McManus 2008, Mabry et al. 2012, Ünlü et al. 2013). Similarly,

in economically deprived areas where in some cases healthcare services are delivered through mobile clinics, a single person or a team can be in charge of both prioritizing and carrying out a full medical examination of the patients who show up at the clinic in the morning (Gove et al. 1999, Razzak and Kellermann 2002, Molyneux et al. 2006, Stillman and Strong 2008, World Health Organization 2008). In such settings, in addition to the typical decision of which patient to prioritize, there is also the question of how to balance the time spent on triage and the time spent on treatment or stabilization.

The patient triage and prioritization problem, particularly in the case of mass-casualty events, is so complex that mathematical formulations that aim to be highly detailed and realistic representations are not likely to lead to implementable solutions. The difficulty arises not only from the numerous factors such decisions would have to consider—and thus that the mathematical optimization problem would have to incorporate—but also from the fact that the reliable estimation of model parameters would be impossible especially considering the lack of available data. Therefore, our main goal in this article is to develop a stylized formulation that captures the essential features of the problem mentioned above and analyze this formulation so as to provide insights that can be helpful in making decisions in practice. In the last part of the paper, we demonstrate how one can design practical policies using our analytical characterization of the optimal policy as well as prior work in mass-casualty triage.

Our model can broadly be described as follows: there are some finite number of patients all of whom are in urgent need of attention from a single medical provider (e.g., a paramedic, nurse, or physician). While all patients are in critical condition, some are in more serious condition than the others and thus need to be served more urgently. In attending to the patients, the provider has various options available. She does not know which patients are in more serious condition. So, she can randomly choose a patient and serve. Alternatively, for each patient she can choose to spend some time on triage to determine the triage class, and thus the urgency level, of the patient. Once a classification is made, she can continue with the service of the patient or move on to another patient who may or may not have already been classified. In parallel with most existing mass-casualty triage protocols such as START (Lerner et al. 2008), which put critical patients in one of two classes, we assume that there are two triage classes named *immediate* (patients with severe and immediately life-threatening injuries) and *delayed* (patients with severe but not immediately life-threatening injuries). Our objective is to determine the actions the provider should take depending on the number and the composition of the patients waiting for attention.

A key issue when formulating this problem is deciding on what the objective function should be. For mass-casualty events, the objective of maximizing the number of survivors is largely accepted in practice, but the question is how exactly that objective can be appropriately captured in a mathematical formulation without rendering it analytically intractable. We discuss our modeling approach in detail in Section 3, but here it suffices to state that our approach mainly rests on the idea that the decline in a patients' survival probability with the passage of time without treatment can be seen as the "waiting cost" for that patient, and thus the minimization of the expected total waiting cost can be interpreted as the minimization of the expected number of deaths. It is also important to note that for our analytical results, we assume that for each patient, the system incurs a delay cost that depends on the triage class of the patient and increases linearly with time. In fact, the only work available to date on survival probabilities for trauma patients (Sacco et al. 2005, 2007; Navin et al. 2009) strongly suggests that survival probabilities do not decrease linearly with time. Nevertheless, as we demonstrate in Section 6 of the paper, our analysis based on this assumption can be used to construct policies that perform well even under realistic conditions, where the assumption is violated.

We review the relevant literature in Section 2 and formally describe our model in Section 3. In Section 4, we provide a complete characterization of the optimal dynamic policy, which allows making decisions based on up-to-date system state—i.e., the number of untriaged patients and the numbers of patients already classified as immediate and delayed all waiting to receive treatment. Under conditions that are most likely to hold in practice, we show that whenever triage identifies an immediate patient, that patient should be served right away; otherwise, the patient should wait until there are no more unclassified or immediate patients. This finding, which essentially deals with the question of how to prioritize when patients are already classified, is not surprising and consistent with the extensive literature that establish the optimality of the $c\mu$-rule under a variety of settings. The more interesting question, which is also the focus of this paper, is when to perform and when to skip triage. As it turns out, this decision depends on the number of unclassified patients and the number of patients classified as delayed. In particular, we find that there is a switching curve that separates the states in which triage should be performed from the others, and we provide a closed-form expression for the curve. One interesting insight that comes out of this characterization is that spending time on triage helps if there are sufficiently many patients but not when there are relatively few. Being overwhelmed with the volume of patients in need of treatment, there could be a temptation to skip triage

and quickly move on to more detailed examination and treatment of the patients in the hopes of saving time. However, our results indicate that this could be a short-sighted decision.

In Section 5, we devise two policies that are simpler alternatives to the optimal policy. The two policies are the *No-Triage* policy, which serves all patients in random order without spending any time in triage, and the *Triage-Prioritize-Class*-1 policy, which performs triage on all patients but serves the triaged patient right away if the patient is classified as class 1—i.e., immediate. We identify conditions under which one is superior to the other, and use both policies as benchmark heuristics later in our computational study.

In Section 6, we show how our analytical results can help devise policies that are likely to perform well in practice. To do that, we first introduce a new mathematical model, which explicitly considers the possibility that the patients might die waiting for treatment, and survival probability functions are chosen in line with the work of Sacco et al. (2005, 2007) and Navin et al. (2009). Then, we describe how our analytical results can be used to construct heuristic policies for this more realistic setting and report the results of a numerical study in which we found that some of these policies perform consistently well across different scenarios. Finally, in Section 7, we provide our concluding remarks and point to some future research directions. Proofs of all of the analytical results are provided in the online supplement.

2. Discussion of the Relevant Literature

Our model and analysis are closely related to the classical job-scheduling literature where *jobs* in our context can be seen as the *patients*, and *servers* or *processors* can be seen as the *medical providers* (e.g., paramedics, nurses, or physicians) who provide triage and treatment services on the scene. A simplified version of our formulation in which class identities of all jobs are known has been studied extensively in the literature. Specifically, when jobs incur linear waiting costs, and the cost rate and the expected service time of class i jobs are respectively given by c_i and $1/\mu_i$, the optimal policy under a variety of conditions is the well-known $c\mu$-rule: a job of class j has priority over a job of class k if and only if $c_j\mu_j > c_k\mu_k$. Starting with Smith (1956), this body of work includes Cox and Smith (1961), Klimov (1974), Harrison (1975), Pinedo (1983), Nain (1989), Argon and Ziya (2009), and Budhiraja et al. (2014), among others. Under convex delay costs, the asymptotic optimality of a generalized version of the $c\mu$-rule, called $Gc\mu$-rule, was established by Van Mieghem (1995) and further studied by Mandelbaum and Stolyar (2004). Our work mainly differs from these articles in that we assume that the class identity of a job is initially unknown and can only be determined through a process called triage, which keeps the server occupied for a certain period of time.

Some of the recent work has considered the job-scheduling problem within the context of mass-casualty triage by either considering models where jobs may renege (patients dying) while waiting or considering time-dependent reward functions, which correspond to time-dependent survival probabilities. Specifically, Argon et al. (2008) consider a single-server two-class model where patients renege from the system with exponential rates that depend on the patient's triage class. They provide a partial characterization of the optimal policy and propose heuristic methods. Uzun Jacobson et al. (2012) consider a more general formulation in which the "reward" obtained through service, which can be seen as the probability that the service will be successful, depends on patient class (though not the time of the service). The authors provide partial characterizations of the optimal policy when there is a single server and propose heuristic methods that can be used even in multiple-server settings. Mills et al. (2013) consider a deterministic fluid model in which there is no reneging but the "reward" for service (survival probability after service) changes with time. Under some realistic conditions, the paper provides a mathematical characterization of the optimal policy and then uses it to propose a prioritization policy that can be implemented in practice.

A number of articles in the literature (e.g., Shumsky and Pinker 2003, Wang et al. 2010, Alizamir et al. 2012, Dobson and Sainathan 2011, Dobson et al. 2013) investigate diagnostic systems that, similar to the triage in our formulation, include a process that reveals some information about the jobs based on which further action is taken. Shumsky and Pinker (2003) consider a two-level service system where the first level acts as a gatekeeper, who first makes an initial diagnosis on arriving customers and then depending on this diagnosis may or may not refer the customers to a specialist. Differing significantly from our focus in this paper, the main objective of Shumsky and Pinker (2003) is to design an incentive mechanism that helps overcome the information asymmetry caused by the gatekeeper being the sole observer of the complexity of the job that each customer presents and the gatekeeper's own treatment ability. One similarity with our work is that, in the model of Shumsky and Pinker (2003), just as in our model, there are two levels of service (triage or serve without triage). However, unlike Shumsky and Pinker (2003), in our formulation, a single server is in charge of both levels of service, and the first level of service (triage) is not mandatory. Wang et al. (2010) study a model where patients may or may not choose to call a diagnostic service center depending on their expectation on the diagnostic accuracy and waiting time. The authors investigate how capacity (staffing) and diagnostic quality decisions should be made.

Alizamir et al. (2012) consider a model where a single server classifies each arriving customer into one of two classes based on the results of a series of independent tests. If the classification is correct, the server receives a reward; otherwise, there is a penalty. Customers who find the server busy join a queue and incurs a waiting cost during their stay in the system. By performing more tests, the server increases the likelihood that a correct classification will be made; however, this increases the waiting time of the customers in the queue. The objective of the paper is to dynamically determine the number of tests to be carried out based on the system state. The fundamental difference between our model and that of Alizamir et al. (2012) is that in our model, the diagnostic process is assumed to be simpler as it consists of a single test and thus the number of tests is not a decision variable. However, unlike Alizamir et al. (2012), we explicitly model the service process that comes after classification and capture the trade-off between the time spent on service and time spent on diagnosis.

Dobson and Sainathan (2011) compare two models: the base model and the prioritized model. In the prioritized model, jobs are first sorted by a pool of homogeneous sorters and then served by another pool of homogeneous processors while there is no sorting in the base model. The authors compare the optimal waiting cost of the prioritized model with that of the base model and identify conditions under which prioritization is beneficial. The work of Dobson and Sainathan (2011) is close to ours in that it also aims to study the trade-off between service capacity allocated to classification and actual service. However, while Dobson and Sainathan (2011) are interested in optimal static design questions by comparing two alternative systems in a multiple-server setting, our goal is to investigate and characterize optimal dynamic decisions for a single-server system. Dobson et al. (2013) study a model in which an investigator collects information from a new customer to decide what work needs to be done in the second step by another server. Once the second step is finished, the customer joins another queue to receive service from the investigator again and then leaves the system. The investigator needs to prioritize between the old and new customers. As we describe in the following section, the models they study are also significantly different from the one we analyze in this paper.

A stream of papers in organizational learning and knowledge management (e.g., March 1991, Gupta et al. 2006, Posen and Levinthal 2012) study what is commonly referred to as "the exploration versus exploitation problem," in which, somewhat similar to our formulation, the main question centers around the allocation of resources to the exploration of new knowledge and the exploitation of existing knowledge. However, unlike in our model, in these papers, exploration typically does not cause delay in exploitation and the two can proceed simultaneously. Furthermore, to the best of our knowledge, none of these papers considers the specific setting we consider in our paper, in which patients are classified into two groups, and their findings do not have any direct implications for our work.

Finally, our work is relevant to a series of papers (e.g., Güneş and Akşin 2004, Gurvich et al. 2009, Armony and Gurvich 2010) that study cross-selling within the context of call centers. As in the case of triage and prioritization we consider in this paper, cross-selling also requires careful balancing of time spent on cross-selling and time spent on actual service. However, the main decision in cross-selling involves when and which customers should be extended offers and, unlike in our case, does not generate information that can be used for service prioritization.

3. The Model

Before we present our mathematical model in detail, we first provide a short discussion on some of the important features of the mass-casualty patient triage problem, explain to what extent the proposed model will successfully capture these features, and give an overview of how the analysis of this model will be used despite its limitations.

3.1. The Mass-Casualty Triage Problem and Our Modeling Approach

A widely accepted utilitarian objective in case of mass-casualty events is to maximize the number of survivors. Obviously, passage of time without treatment has a negative effect on the survival chances of each patient, and thus it makes sense to talk about the "cost" of waiting. But what exactly is this cost? To understand this, it is very important that we distinguish between the two different ways that waiting can affect a patient's probability of survival. First, one consequence of a patient waiting for treatment could simply be that the patient might die by the time it is that patient's turn for treatment. Clearly, the longer the patient waits, the higher the chance the patient will die before treatment. However, even if the patient is alive by that time, the probability of survival is not the same as it would be without waiting. The overall deterioration of the patient as a result of waiting decreases the chances of a successful operation and the eventual survival of the patient. This is the second way that waiting affects a patient's probability of survival. While the survival probability decreases with time either way, from a modeling point of view, one important difference is that if a patient dies before treatment, that patient no longer needs service.

Thus, to fully capture the effect of waiting on the patients, it would be reasonable to consider a model where (i) each patient has a *remaining life time* (patience time) at the end of which the patient dies (abandons the

system) if she or he is not provided the necessary treatment by that time, and (ii) if the patient is alive when this patient's turn for treatment comes, she or he dies with some probability that increases with the waiting time of the patient. Assuming that each life lost (either before or after treatment) would have a cost of one unit, the objective would be to minimize the expected number of deaths or equivalently the expected total cost. While developing such a model is straightforward, its analysis is extremely difficult partially due to the fact that even under some restrictive assumptions such as deterministic triage and service times, which help incorporate the passage of time in the system state in a relatively convenient way, the resulting transition probability structure is too dense to permit clean analytical characterizations (see the formulation in Section 6.1). In fact, as one can see from the analysis of Uzun Jacobson et al. (2012), where the impact of waiting is captured through abandonments alone (ignoring the effect of the passage of time on the success of service) and the only question is how to prioritize patients who have already gone through triage, the optimal policy has a complex structure, which only permits highly limited analytical characterizations. Therefore, in this paper, we follow an alternative approach according to which we simplify the formulation in a way that makes mathematical analysis possible but then investigate whether the results of this study would be useful in practice by using the more complex and realistic formulation as a test bed.

Specifically, in our mathematical model, we assume that the system incurs a fixed cost for each unit of time a patient waits. This fixed per-unit time cost depends on the triage class of the patient, but it does not change with time. Furthermore, patients do not renege from the system while waiting for their treatment. In this model, waiting cost can be seen as capturing the two different ways waiting impacts patient survival as explained above. Clearly, this would be an approximation not only because studies suggest that survival probabilities are not linear functions of time, but also because the problem would be structurally different with and without reneging. Nevertheless, even though our mathematical model ignores the possibility of reneging, this does not mean the fact that some of the patients died while waiting would be ignored when it comes to practical implementation of the heuristic methods that are based on the analysis of this model. (In other words, the methods would wisely not suggest treatment of dead patients.) And the nonlinearity of the survival probability functions might possibly be overcome by using linear approximations. Therefore, the analysis of this simplified model has the potential to lead to methods that perform well. With this motivation, we next describe our model in detail and present our analysis. Later, in Section 6.4, we provide the results of a detailed numerical investigation, which shows that our analysis indeed leads to heuristic methods that perform well even when the linear waiting cost assumption is relaxed and patients may possibly die and thus renege from the system while waiting.

3.2. Model Description

We consider a scenario in which an unexpected event triggers the sudden appearance of a number of patients in need of treatment. More specifically, we assume that at time $t = 0$, there are $N \geq 2$ patients waiting for treatment. For reasons that will be clear shortly, we refer to these patients as class 0 patients. There will be no new patient arrivals. There is a single provider, which we will refer to as the *server* for expositional convenience, and the treatment this server provides to the patients is referred to as the *service*.

The server does not have to triage the patients to serve them. In other words, each patient can be served as a class 0 patient. However, she can choose to perform triage, at the end of which the patient is put in one of two classes, class 1 or class 2. Following the terminology of the widely adopted mass-casualty triage protocol START, class 1 patients can be seen as *immediate* patients and class 2 patients can be seen as *delayed* patients. We let $\alpha_i \geq 0$ for $i = 1, 2$ denote the probability that a class 0 patient is classified as class i as a result of triage, thus $\alpha_1 + \alpha_2 = 1$. Once a patient is classified, the server either serves the patient immediately or delays the service of the patient temporarily, making note of the patient's class information, and moves on to another patient. Once the service of the patient is over, she or he leaves the system.

Let $f_i(t)$ denote the expected cost incurred if a class i patient spends t time units in the system, $i = 0, 1, 2$. For our mathematical analysis, we assume that $df_i(t)/dt = r_i \geq 0$ for $t \geq 0$ and $i = 0, 1, 2$, which means that for each unit of time a class i patient spends waiting, in triage, or in service, the system incurs an expected cost of r_i. We relax this linear cost assumption later in our numerical study. Let τ_i denote the expected service time for a class i patient, and c_i denote the total expected cost a class i patient will incur while receiving service. Note that we do *not* assume that $c_i = r_i\tau_i$ so that we allow the service time and the waiting cost rate for a random class i patient to possibly depend on each other. Triage times are assumed to be independent of the service times and patients' class identities. This is a reasonable assumption for systems where, as in the case of mass-casualty triage and prioritization, there is a predetermined procedure to be used for classification of the patients. We use u to denote the expected time it takes to triage one patient. The objective is to minimize the total expected cost of all of the patients. Throughout the paper, we assume that the following two conditions hold:

Assumption 1. (i) $0 \leq \tilde{\tau} = \tau_0 - \alpha_1\tau_1 - \alpha_2\tau_2 < u$; (ii) $\tilde{c} = c_0 - \alpha_1 c_1 - \alpha_2 c_2 < r_0 u$.

To understand what exactly these conditions imply, consider a single class 0 patient in need of service. The first inequality of Assumption 1(i) implies that knowing a patient's triage class helps reduce the patient's expected service time, but the second inequality implies that it takes longer for the server to first triage the patient and serve afterward than to serve the patient right away without triage. In other words, if the class information were readily available, that would help reduce the expected time it would take to serve the patient, but if triage is needed to obtain the class information, the total expected time spent for the patient would be longer. Assumption 1(ii) implies that it is also more costly for the server to first triage the patient and serve afterward than to serve the patient right away without triage. This means that, for a single patient in isolation (i.e., when $N = 1$), triage has no benefit. This is a realistic assumption in settings like mass-casualty triage and prioritization, where triage merely serves as a sorting mechanism and does not involve any specialized preprocessing that would somehow reduce the total triage plus service time or waiting cost for any individual patient. (In our analysis, Assumption 1 helps in coming up with a clear characterization of the optimal choice between performing triage and serving a class 0 patient without triage. Neither of the two conditions of Assumption 1 by itself without the other is sufficient to determine the optimal action. However, if neither of them holds, we can show that the optimal decision is to always triage class 0 patients.)

We also assume, without loss of generality, that $r_1/\tau_1 \geq r_2/\tau_2$. Thus, the well-known $c\mu$-rule implies that if all patients are already classified as class 1 or class 2, the optimal decision is to prioritize class 1 patients. Note that prioritization of a single class over the other (the immediate patients over the delayed patients) is also consistent with the widely adopted mass-casualty triage protocol START. It is important to note that the assumption $r_1/\tau_1 \geq r_2/\tau_2$ together with Assumption 1 also imply that $r_1/\tau_1 \geq r_0/\tau_0$. Therefore, in our analysis, it will be sufficient to consider two cases: $r_0/\tau_0 \geq r_2/\tau_2$ and $r_0/\tau_0 < r_2/\tau_2$.

Our problem can be formulated as a Markov decision process where the decision epochs are time zero and triage and service completion times. (We assume that service is non-preemptive.) The state of the system can then be denoted by the triplet (i, k_1, k_2), where i represents the number of class 0 patients, and k_1 and k_2 denote the number of patients that have been classified as class 1 and class 2 but not yet served, respectively. Since we have N patients in total, the state space can be described as $\mathcal{S} = \{(i, k_1, k_2): i, k_1, k_2 \geq 0, i + k_1 + k_2 \leq N\}$.

Using a sample-path argument, it is straightforward to show that keeping the server idle is suboptimal. This allows us to ignore idling as an admissible action. Then, in a given state $s = (i, k_1, k_2)$, the available actions for the server are as follows: *SU*, serve a class 0 patient without triage (only available if $i \geq 1$); *Tr*, triage a class 0 patient (only available if $i \geq 1$); *SC1*, serve a class 1 patient (only available if $k_1 \geq 1$); and *SC2*, serve a class 2 patient (only available if $k_2 \geq 1$). In general, it is possible that there is more than one optimal action for any given state. If that is the case, we choose the action that is listed earlier in the action set {*SC1*, *SU*, *Tr*, *SC2*}. For instance, *SC1* has precedence over all of the other actions. While this assumption is not crucial, it allows us to ensure that there is a unique optimal policy, which in turn simplifies the presentation of the results.

We define $a^*(s)$ for $s \in \mathcal{S}$ to be the optimal action in state s. We also let $V_\pi(i, k_1, k_2)$ denote the total expected cost under policy π, and $V(i, k_1, k_2) = \min_\pi\{V_\pi(i, k_1, k_2)\}$ be the total expected cost under an optimal policy starting from state (i, k_1, k_2) with no service or triage in progress. We can write the optimality equations as follows:

$$
\begin{aligned}
&V(i, k_1, k_2)\\
&\quad = \min\{\alpha_1 V(i-1, k_1+1, k_2) + \alpha_2 V(i-1, k_1, k_2+1)\\
&\qquad + (ir_0 + k_1 r_1 + k_2 r_2)u, V(i-1, k_1, k_2) + c_0\\
&\qquad + [(i-1)r_0 + k_1 r_1 + k_2 r_2]\tau_0, V(i, k_1-1, k_2) + c_1\\
&\qquad + [ir_0 + (k_1-1)r_1 + k_2 r_2]\tau_1, V(i, k_1, k_2-1) + c_2\\
&\qquad + [ir_0 + k_1 r_1 + (k_2-1)r_2]\tau_2\},\\
&\qquad\qquad \forall (i, k_1, k_2) \in S\backslash(0,0,0),\\
&V(0,0,0) = 0, \quad \text{and} \quad V(s) = \infty, \quad \forall s \notin S. \qquad (1)
\end{aligned}
$$

Finally, it is natural to assume that the initial state is $(N, 0, 0)$ so that we start with N class 0 patients (and no service or triage in progress), and consequently the objective is to determine the policy π that minimizes $V_\pi(N, 0, 0)$. However, as it should be clear in our analysis, this assumption does not change our analysis in any way, and the results would go through regardless of the initial state.

4. Complete Characterization of the Optimal Policy

If there was no option to triage and the decision only involved prioritizing among the three classes of patients, we already know from the $c\mu$-rule that patients would be prioritized according to their r_i/τ_i values with higher values of r_i/τ_i indicating higher priorities. For our problem, as we explain in the following, this index ordering is still highly relevant but, not surprisingly, insufficient to fully describe the optimal policy.

To provide a complete characterization of the optimal policy, it will be sufficient to consider two separate cases: (i) $r_0/\tau_0 \geq r_2/\tau_2$; (ii) $r_0/\tau_0 < r_2/\tau_2$. Per the $c\mu$-rule, the ratio r_i/τ_i can be seen as a measure of the relative urgency or importance of class i

patients. class 0 patients are those patients for whom we do not have a clear idea about their urgency. The goal with triage is to gain some information on these patients so that they can be identified as *immediate* (class 1) or *delayed* (class 2). Thus, it would be natural to assume the urgency measure of a class 2 patient to be smaller than the urgency measure of a random patient we do not know anything about—i.e., an untriaged class 0 patient. Therefore, at least for our main motivational purposes, the more practically relevant setting is case (i). We start our analysis with that case. However, we do provide a description for the other case later in this section for completeness.

Theorem 1. *Suppose that* $r_0/\tau_0 \geq r_2/\tau_2$ *and consider state* $(i, k_1, k_2) \in \mathcal{S}$:

(i) *If* $k_1 \geq 1$, *then* $a^*(i, k_1, k_2) = SC1$*—i.e., as soon as the server identifies a class* 1 *patient, that patient should be served next.*

(ii) *If* $i + k_1 \geq 1$, *then* $a^*(i, k_1, k_2) \neq SC2$*—i.e., it is optimal to serve a class* 2 *patient only when there are no class* 0 *or class* 1 *patients.*

(iii) *There exists a linear function* $L(\cdot)$ *such that for any state* $(i, 0, k_2) \in \mathcal{S}$, *where* $i \geq 1$ *and* $k_2 \geq 0$, *if* $k_2 \geq L(i)$, *then* $a^*(i, 0, k_2) = SU$*—i.e., the optimal action is to serve without triage; otherwise,* $a^*(i, 0, k_2) = Tr$*—i.e., the optimal action is to perform triage. Furthermore,*

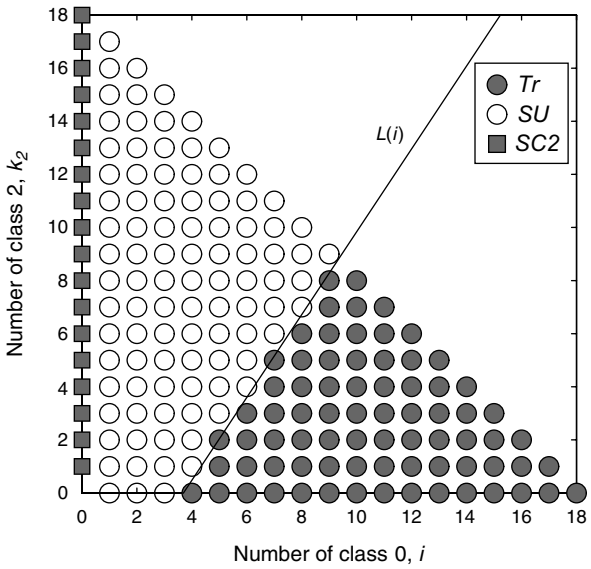
$$L(i) = \frac{r_0(\tilde{u} - u)}{r_2(u - \tilde{\tau})} i - \frac{r_0 \tilde{u} - \tilde{c}}{r_2(u - \tilde{\tau})}, \tag{2}$$

where $\tilde{u} = \alpha_1(r_1\tau_0 - r_0\tau_1)/r_0$ *and* $\tilde{c}$ *and* $\tilde{\tau}$ *are as defined in Assumption* 1.

Parts (i) and (ii) of Theorem 1 clearly delineate the regions where serving patients classified as class 1 and class 2 are optimal. Specifically, *SC1* has precedence over all other actions no matter the current state. This means that as soon as a triage results in identification of a class 1 patient, the next action is to serve that patient. On the other hand, *SC2* is at the bottom of the priority list, meaning that the service of class 2 patients starts at the end when there are no more class 1 or class 0 patients waiting.

Part (iii) of Theorem 1 describes the optimal action when there are no class 1 patients (i.e., $k_1 = 0$) but there is at least one class 0 patient (i.e., $i \geq 1$). Recall that in such a state, the server can choose to either triage or directly serve a class 0 patient. (We know from part (ii) of the theorem that serving a class 2 patient, if there is one, is suboptimal.) It turns out that whether or not doing triage is optimal depends on the system state. More specifically, there is a line that separates the states in which doing triage is optimal from the states in which serving without triage is optimal. In Theorem 1(iii), we not only prove this structural property of the optimal policy, but also provide a closed-form expression for this line. See Figure 1 for a visual

Figure 1. Visual Description of the Optimal Policy When $k_1 = 0$ and $N = 18$, $\alpha_1 = 0.2$, $u = 0.5$, $r_0 = 8.4$, $r_1 = 10$, $r_2 = 2$, $\tau_0 = 2.4$, $\tau_1 = 2$, $\tau_2 = 4$, $c_0 = 17.6$, $c_1 = 20$, $c_2 = 8$



demonstration of the optimal policy structure for a specific example.

Theorem 1 provides interesting insights into the decision of when to do triage and when to skip it. Suppose that initially at time zero there are some N class 0 patients and no class 1 or class 2 patients, as that would be the case in the immediate aftermath of a mass-casualty event just before the start of patient triage and treatment. This means that at time zero, in Figure 1, the system starts on the x axis at $i = N$. If N is large, meaning that there are too many patients waiting to be served and we have no information regarding which ones are more important, one might be tempted to skip triage since performing triage will further lengthen the waiting times, which are already likely to be too long. With too many patients to serve, spending time on triage might seem like an unwise use of time. In contrast, when N is small, triage might not seem all that harmful since waiting times are not going to be too long even with triage. As we explain in the following, however, this reasoning is flawed.

Theorem 1 states that—as one can also easily verify referring to Figure 1—when the number of class 0 patients is sufficiently large (initially more than or equal to four for the example whose solution is depicted in the figure), it is optimal to start with triage and continue to do so as long as the number of class 0 patients and the number of class 2 patients keep the state space under the line. (Note that if a class 1 patient is identified, that patient is served right away.) Once the threshold line is passed, the optimal policy starts serving patients without triage until there are no more class 0 patients waiting. class 2 patients, who would

have been identified as such earlier, are served at the end. If the number of class 0 patients is small (initially less than four in the example), then the optimal policy is simply to serve all of the patients without triage. Thus, contrary to the argument above, precisely because there are too many patients, one cannot afford to skip triage. Even if triage is skipped, service will take quite a long time anyway. Therefore, it makes sense to spend some time at the beginning (specifically as long as the system state is to the right of the threshold line) to perform triage in an effort to at least prevent the waiting times for *immediate* patients from getting too long. On the other hand, when there are few patients, service of all patients, regardless of how urgent their conditions are, will not take too much time. Therefore, the value of class information that will be obtained through triage does not justify the additional waiting that all patients will have to endure.

It is also interesting to note that the optimal policy appears to prefer performing triage when the expected fraction of class 1 patients is sufficiently high. To see that, first note that $i + k_2$ is the total number of unserved patients in the system when the system state is $(i, 0, k_2)$ and the expected fraction of class 1 patients is $\alpha_1 i/(i + k_2)$. When i is large in comparison with k_2, this fraction is large and is close to α_1, and triage is the preferred option. However, when the fraction is small, then the optimal policy chooses to skip triage.

We now consider the opposite case, where $r_0/\tau_0 < r_2/\tau_2$. As we discussed above, this case is of somewhat less practical interest since we view triage as a procedure that merely helps in obtaining information on the patients and sorting them out with respect to their relative urgency. Nevertheless, analysis under this condition might still be of interest if what we call triage is interpreted as some sort of preprocessing that results in such a change in the urgency measure of the patients.

Theorem 2. *Suppose that $r_0/\tau_0 < r_2/\tau_2$. Then, there exists an optimal policy under which* (i) *no patient goes through triage*; (ii) *patients are served in accordance with the $c\mu$-rule—i.e., a patient with a higher value of r_i/τ_i, $i = 0, 1, 2$ gets a higher priority.*

Theorem 2 essentially says that under the condition we stated above, triage has no benefit and the prioritization policy should simply follow the $c\mu$-rule. One implication of this result is that if at time zero there are N patients, none of which are triaged, then it is optimal to not perform triage on any one of the patients and serve them all without triage.

5. Simpler Alternatives to the Optimal Policy

In the previous section, we provided a complete characterization of the optimal policy. While the optimal policy is relatively simple, it is possible to devise even simpler policies, which may not be optimal but would perform well under certain conditions. Such policies may be preferred over the optimal policies because of their ease of implementation in practice. In this section, we will investigate some of these simpler alternatives, some of which will also serve as benchmark policies in our computational study.

For ease of exposition, we assume in this section that at time zero, all N patients in the system are from class 0. From Theorem 2, we already know that the policy of not triaging any patient is in fact the optimal policy when $r_0/\tau_0 < r_2/\tau_2$. Therefore, we focus on the case where $r_0/\tau_0 \geq r_2/\tau_2$.

An obvious candidate for a simple policy is to not triage any of the patients and serve them in random order. Another possibility is to triage all of the patients regardless of the system state. In this case, however, one needs to specify when and how exactly triage information will be used. One can first complete triage of all of the patients and then move on to the service of the patients. In accordance with the $c\mu$-rule, class 1 patients would have priority over class 2 patients. Alternatively, if triage identifies a patient from a particular class, the server serves that patient before moving onto the triage of the rest of the patients. Patients from the other class are served once the triage of all of the patients is complete. In this case, intuitively it would make sense to give priority to class 1 patients, but in fact there are examples that show that it is not always better than prioritizing class 2 patients. Therefore, it would be reasonable to consider the policy that prioritizes class 2 as well. (Note that it is easy to show that there is no benefit to be gained from the triage of the last unclassified patient regardless of which class has priority and whether or not the service is delayed until all triage is complete. Therefore, in what follows "triage of all patients" means "triage of all patients except the very last untriaged patient.")

The following proposition helps eliminate some of the potential policies described above for further consideration as they can be shown to be inferior to the others.

Proposition 1. (i) *If all patients have to go through triage, it is strictly better for the server to serve class* 1 *patients as soon as they are identified than to complete triage of all patients first and then move on to the service of all of the classified patients.*

(ii) *It is strictly better for the server to skip triage and serve patients in random order than to triage all of the patients, serve class* 2 *patients as soon as they are identified, and serve class* 1 *patients at the end.*

Proposition 1(i) simply says that delaying the start of service until every single patient is classified does not work well. This is because, once a patient that has a

high priority is identified, there is no point in delaying the service of that patient. We know for sure that no other patient will get a higher priority. Part (ii) of the proposition says that skipping triage altogether and serving patients in a random order always works better than triaging patients while serving class 2 patients as soon as they are identified. Interestingly, there are examples that show that it might be better to prioritize class 2 patients over class 1 patients, but that can only happen if skipping triage altogether and serving patients in random order is superior to any prioritization policy with triage. Thus, we can focus our attention to the following two simple policies:

No-Triage Policy (NT): Patients are served in random order. No patient goes through triage.

*Triage-Prioritize-Class-*1 *Policy* (TP1): Each patient, with the exception of the last one, goes through triage in random order. If a patient is classified as class 1, she or he is served right away; otherwise, the patient is put aside to be served later. When the triage of $N-1$ patients is completed, the remaining untriaged patient is served followed by all class 2 patients.

We denote the total expected cost under policies NT and TP1 by V_{NT} and V_{TP1}, respectively. Because of the relatively simple structure of the two policies, we can come up with closed-form expressions for V_{NT} and V_{TP1}. We refer the reader to the online supplement for the expressions as well as their derivations. The following proposition identifies the conditions under which one policy is superior to the other.

Proposition 2. *Suppose that $r_0/\tau_0 \geq r_2/\tau_2$ and the initial system state is $(N,0,0)$. Then, $V_{\mathrm{TP1}} \leq V_{\mathrm{NT}}$ if and only if $u \leq \beta$ where*

$$\beta = \max\left\{\frac{(N/2)\alpha_2(r_0\tau_2 - r_2\tau_0) + \tilde{c} - r_0\tilde{\tau}}{(N/2)(\alpha_2 r_2 + r_0) + \alpha_1 r_1} + \tilde{\tau}, 0\right\}. \quad (3)$$

Proposition 2 confirms the intuition that when the expected triage time is sufficiently short—i.e., the service provider does not need to spend a long time to obtain information and classify class 0 patients—the benefit obtained through triage could offset the additional cost incurred as a result of triage, and TP1 outperforms NT. More specifically, the proposition gives a precise description of what we mean by the triage time being *sufficiently short*.

One important question is whether there are certain conditions under which either TP1 or NT is in fact optimal. When $r_0/\tau_0 < r_2/\tau_2$, we know that NT is optimal, but how about when $r_0/\tau_0 \geq r_2/\tau_2$? It would be natural to expect that when the expected triage time is sufficiently short (it might help to think of the limiting case where it is zero), it would be optimal for all patients to go through triage, and conversely when the expected triage time is sufficiently long, it would be optimal for none of the patients to go through triage. Indeed, we can prove that is the case. The following proposition formalizes this result and clearly describes what would qualify as sufficiently short and what would qualify as sufficiently long. Let

$$u_1 = \min\left\{\tilde{u}, \tilde{u} - \frac{r_0\tilde{u} - \tilde{c}}{Nr_0}\right\},$$
$$u_2 = \min\left\{\frac{r_0\tilde{u} + \tilde{c}}{2r_0}, \frac{r_0\tilde{u} + \tilde{c} + (N-2)r_2\tilde{\tau}}{2r_0 + (N-2)r_2}\right\}.$$

Proposition 3. *Suppose that $r_0/\tau_0 \geq r_2/\tau_2$ and the initial system state is $(N,0,0)$. Then,*

(i) *policy* NT *is optimal if and only if $u \geq u_1$;*

(ii) *policy* TP1 *is optimal if and only if $u \leq u_2$;*

(iii) *furthermore, u_1 is nondecreasing and u_2 is nonincreasing in N.*

When the expected triage time is as long as described in Proposition 3(i), the information that one would get through triage is simply not worth it. Hence, the optimal policy is to serve all of the patients directly without triage. When the expected triage time is as short as described in Proposition 3(ii), one can "afford" to triage all of the patients; however, in line with Theorem 1, if a class 1 patient is identified as a result of triage, that patient should be served first before moving on to the triage of the remaining patients. When the expected triage time is between u_1 and u_2, then neither NT nor TP1 is optimal. The optimal policy is state dependent as characterized by Theorem 1. Note that Proposition 3 can be seen as a strengthened version of Proposition 2 since the former provides necessary and sufficient conditions for TP1 and NT to be optimal, whereas the latter delineates the region where one performs better than the other. While neither of the results implies the other, the two are in agreement (as expected) on the relationship between the expected triage time and the performances of the two policies.

Part (iii) of Proposition 3 provides an interesting insight into the effect of N, the initial number of patients, on the optimal policy. We can see that as N increases, the parameter region in which NT is optimal and the parameter region in which TP1 is optimal both shrink (or at least they do not get larger). This suggests that simple policies like NT and TP1 are more likely to be good choices when there are relatively few patients initially in the system.

6. Nonlinear Waiting Costs with Reneging Patients: A Numerical Study on Mass-Casualty Triage

Our analysis so far in this paper has been based on two crucial assumptions that may be questionable in the context of patient triage particularly in the case of mass-casualty events. The first assumption is that the "waiting cost" that the system incurs for each patient is a linear function of the patient's waiting time.

In mass-casualty patient triage, the waiting cost of a patient can be seen as the decline in the probability that the patient will survive the service (operation) she or he will have to go through, assuming that the patient is still alive by the time of the service. With this interpretation, the linear waiting cost assumption may not adequately capture the reality. The second assumption is that no patient dies (or reneges) while waiting for her or his turn for the service, which might not be true as well. The objective of this section is to investigate whether we can use our analysis (more specifically, our optimal policy characterization) to develop policies that can be used under more realistic conditions, where waiting costs (changes in survival probabilities) are not linear in time and patients might die while waiting.

In the following, we first introduce the mathematical framework we will use to investigate the performances of the policies we will be proposing. This more realistic (at least in certain respects) framework still needs to abide by certain assumptions so that the "optimal" policy can be computed and therefore the performances of our policies properly assessed. Then, we describe how we can use our analytical results, more specifically Theorem 1, to develop heuristic methods and devise three new policies. It is important to note that these policies are not custom designed for the mathematical framework we will be introducing and can be easily implemented as long as some key model parameters are properly estimated. We then describe the specific mass-casualty scenario we consider in the numerical study and present our findings regarding how these policies perform in comparison with the optimal policy and some of the benchmark policies.

6.1. Description of the Model with Nonlinear Waiting Costs and Reneging

Let X_i denote the lifetime (time until reneging) without treatment for a random class i patient and assume that for $i = 1, 2$, X_i is an independent random variable with $G_i(t) \equiv P\{X_i \le t\}$. If a patient's lifetime ends before the patient is taken into service, then the patient reneges, she or he no longer needs service, and the system incurs a cost of one unit. Let $\alpha_i(t)$ for $i = 1, 2$ denote the probability of labeling a random class 0 patient as class i if the patient goes through triage at time t. Note that this probability is time dependent (unlike the model in Section 3) because the remaining lifetime distributions for class 1 and class 2 patients are different. By letting $\alpha_i = \alpha_i(0)$ for $i = 1, 2$ denote the probability that a random class 0 patient who goes through triage at time zero would be classified as class i, we have for any $t \ge 0$,

$$\alpha_i(t) = P\{Z = i \mid X_0 > t\} = \frac{P\{Z = i,\ X_0 > t\}}{P\{X_0 > t\}} = \frac{P\{X_0 > t \mid Z = i\}P\{Z = i\}}{\sum_{i=1}^{2} P\{X_0 > t \mid Z = i\}P\{Z = i\}}, \quad i = 1, 2,$$

where Z denotes the class identity of a random class 0 patient after the patient is triaged. Then,

$$\alpha_1(t) = \frac{P\{X_1 > t \mid Z = 1\}P\{Z = 1\}}{\sum_{i=1}^{2} P\{X_i > t \mid Z = i\}P\{Z = i\}} = \frac{\alpha_1 \bar{G}_1(t)}{\alpha_1 \bar{G}_1(t) + \alpha_2 \bar{G}_2(t)}, \quad \alpha_2(t) = 1 - \alpha_1(t), \tag{4}$$

where $\bar{G}_i(t) \equiv 1 - G_i(t)$. Let $p_i(t, \Delta t)$ denote the probability that a class i patient survives for another Δt time units given that the patient has survived the first t time units. Then,

$$p_i(t, \Delta t) = P\{X_i > t + \Delta t \mid X_i > t\} = \frac{P\{X_i > t + \Delta t\}}{P\{X_i > t\}} = \frac{\bar{G}_i(t + \Delta t)}{\bar{G}_i(t)}, \quad i = 1, 2.$$

$$p_0(t, \Delta t) = \alpha_1(t)p_1(t, \Delta t) + \alpha_2(t)p_2(t, \Delta t) = \frac{\alpha_1 \bar{G}_1(t + \Delta t) + \alpha_2 \bar{G}_2(t + \Delta t)}{\alpha_1 \bar{G}_1(t) + \alpha_2 \bar{G}_2(t)}.$$

We assume that service times and triage times are deterministic, do not depend on the class of the patient, and are denoted by τ and u, respectively. Note that deterministic triage and service times allow us to compute the optimal policy and make comparisons with the performances of our policies.

Recall that in our model described in Section 3, we used $f_i(t)$ to denote the expected cost the system will incur for a class i patient who spends t time units waiting. Here, we assume that each death patient incurs a cost of 1 and thus $f_i(t)$ corresponds to the probability that a class i patient who has survived by time t and is taken into service at time t will not have a successful service and eventually die due to the injuries caused by the mass-casualty event. Let $b(k; n, p)$ denote the probability of getting exactly k successes in n Bernoulli trials each of which yields success with probability p—i.e.,

$$b(k; n, p) = \binom{n}{k} p^k (1 - p)^{n-k}. \tag{5}$$

When $i \ge 1$, $k_1 \ge 1$ and $k_2 \ge 1$, all four possible actions (triage, serve class 0, serve class 1, serve class 2) are available and the optimality equation for this case can be written as

$$\begin{aligned} V(i, k_1, k_2, t) = \min \Bigg\{ & \sum_{i'=0}^{i-1} b(i'; i - 1, p_0(t, u)) \sum_{k_1'=0}^{k_1} b(k_1'; k_1, p_1(t, u)) \\ & \cdot \sum_{k_2'=0}^{k_2} b(k_2'; k_2, p_2(t, u))(\alpha_1(t)V(i', k_1' + 1, k_2', t + u) \\ & + \alpha_2(t)V(i', k_1', k_2' + 1, t + u) \\ & + (i + k_1 + k_2 - i' - k_1' - k_2' - 1)), \end{aligned}$$

$$\cdot \sum_{i'=0}^{i-1} b(i'; i-1, p_0(t,\tau)) \sum_{k_1'=0}^{k_1} b(k_1'; k_1, p_1(t,\tau))$$
$$\cdot \sum_{k_2'=0}^{k_2} b(k_2'; k_2, p_2(t,\tau))(V(i', k_1', k_2', t+\tau) + (i + k_1 + k_2 - i' - k_1' - k_2' - 1 + f_0(t))),$$
$$\cdot \sum_{i'=0}^{i} b(i'; i, p_0(t,\tau)) \sum_{k_1'=0}^{k_1-1} b(k_1'; k_1 - 1, p_1(t,\tau))$$
$$\cdot \sum_{k_2'=0}^{k_2} b(k_2'; k_2, p_2(t,\tau))(V(i', k_1', k_2', t+\tau) + (i + k_1 + k_2 - i' - k_1' - k_2' - 1 + f_1(t))),$$
$$\cdot \sum_{i'=0}^{i} b(i'; i, p_0(t,\tau)) \sum_{k_1'=0}^{k_1} b(k_1'; k_1, p_1(t,\tau))$$
$$\cdot \sum_{k_2'=0}^{k_2-1} b(k_2'; k_2 - 1, p_2(t,\tau))(V(i', k_1', k_2', t+\tau) + (i + k_1 + k_2 - i' - k_1' - k_2' - 1 + f_2(t)))\Big\}, \quad (6)$$

where the four terms inside of the minimum from the first to the last, respectively, corresponds to the actions *triage*, *serve class* 0, *serve class* 1, and *serve class* 2. For all other states, where at least one of i, k_1, or k_2 is zero, the optimality equations can similarly be written.

6.2. Heuristic Policies

We propose three different policies, all based on our analytical results provided in Section 4. For all three policies, we first fit least-squares lines to the death probability functions $f_i(t)$ (which correspond to the cost functions in Section 3) for the immediate and delayed patients. When fitting the least-squares lines, we assume that the cost function is defined over the interval $[t_0, t_0 + \max_{i=0,1,2} N(\tau_i + u)]$, where t_0 is the time when the response effort starts and $t_0 + \max_{i=0,1,2} N(\tau_i + u)$ is the maximum expected time by which all of the patients in the system are served and $i = 0,1,2$ respectively corresponds to unclassified, immediate, and delayed patients. All three policies rely on the idea of using these least-squares lines as approximations for the actual cost functions and making use of the analytical characterizations of the optimal policy under the assumption of linear delay costs (Theorems 1 and 2).

(i) *Dynamic Threshold Policy* (DTP): For any given state, this policy prescribes taking the action that is optimal under the assumption that waiting costs for the immediate and delayed patients are given by the least-squares lines that are fit to the "actual" waiting cost functions—i.e., death probability functions. Specifically, this policy takes actions in accordance with Theorems 1 and 2, where all of the parameters and the threshold function $L(\cdot)$ are computed using the slopes of the fitted lines in place of the linear cost parameters r_1 and r_2. We call this policy "dynamic threshold policy" because, unlike the other two policies described below, the threshold on the number of unclassified patients, which determines whether or not triage should be performed, changes with the number of patients classified as delayed.

(ii) *Static Threshold Policy* 1 (STP-1): Similar to DTP, this policy also bases its actions on Theorems 1 and 2 assuming linear costs with slopes given by the slopes of the least-squares lines. However, the only exception is that this policy uses a static threshold value on the number of unclassified patients to determine whether or not triage should be carried out. Specifically, L_1, the threshold for the policy STP-1, is given by

$$L_1 = \frac{r_0\tilde{u} - \tilde{c} + N r_2(u - \tilde{\tau})}{r_0(\tilde{u} - u) + r_2(u - \tilde{\tau})}. \quad (7)$$

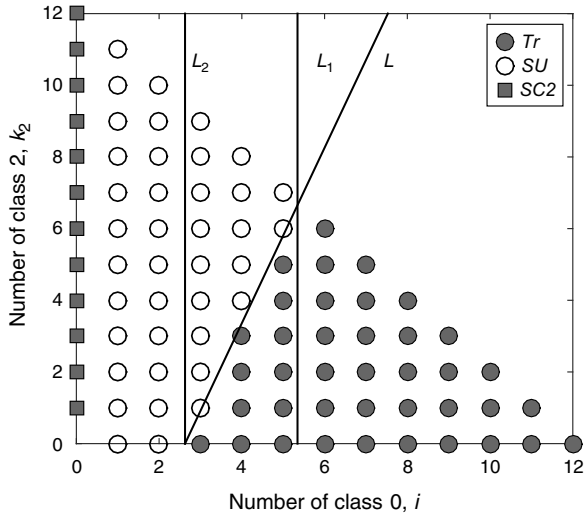
Note that L_1 is not a function of k_2, which means that it does not change with the number of patients classified as delayed.

(iii) *Static Threshold Policy* 2 (STP-2): As in the case of DTP and STP-1, this policy also bases its actions on Theorems 1 and 2 assuming linear costs with slopes given by the slopes of the least-squares lines. The exception is again in the way the threshold value is calculated. Specifically, for STP-2, the threshold L_2 has the expression

$$L_2 = \frac{r_0\tilde{u} - \tilde{c}}{r_0(\tilde{u} - u)}. \quad (8)$$

Figure 2 provides a visual demonstration of how these three policies differ from each other. In the figure, L, which is the threshold line for DTP, directly comes from (2) and depends on the number of unclassified (class 0) and delayed (class 2) patients. The main motivation behind developing policies STP-1 and STP-2 as alternatives to DTP is to investigate whether

Figure 2. Visual Description of the Three Heuristic Policies



simpler policies, which have vertical threshold lines and thereby make decisions based on the number of untriaged patients alone, can also perform well. The threshold line for STP-1, L_1, is defined as the vertical line that passes through the intersection of L and the right edge of the state space described by the line $i + k_2 = N$, and the threshold line for STP-2, L_2, is the vertical line that passes through the x-intercept of L. This is how we obtain the expressions for L_1 and L_2 given by (7) and (8). From Figure 2, we can see that STP-1 triages fewer patients than DTP and STP-2 triages more patients than DTP.

6.3. Description of a Mass-Casualty Scenario

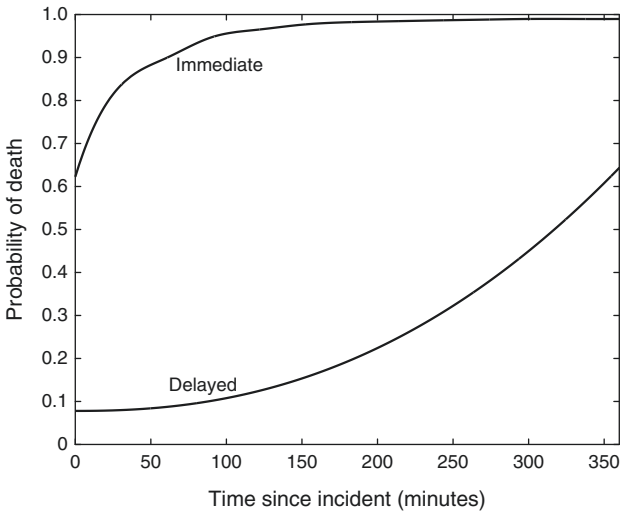
We consider a battlefield mass-casualty scenario in which as a result of an unexpected attack or bombing, a single paramedic is suddenly presented with a number of military-age casualties all in urgent need of some basic on-site treatment for survival until resources are available for transporting them to nearby treatment facilities for higher-level care. While there are different triage systems that are used in responding to these types of events, most put patients into one of four classes typically named as *expectant, immediate, delayed,* and *minimal*. Expectant patients are those who have no chance of survival, and minimal patients are those who do not have any serious life-threatening injuries. Thus, treatment priority is given to immediate and delayed patients, and the success of the response effort is ultimately determined by the way the patients in these two groups are triaged, prioritized, and treated.

While data are typically available for emergency responses to daily events, data in the case of mass-casualty events, particularly in case of triage and treatment in battlefields, are severely limited. To the best of our knowledge, there is no work that investigates how long it takes to triage and treat casualties in such environments. This poses a challenge to testing our policies through a numerical study. To overcome this challenge, at least to the extent possible, we consulted with David A. Masneri, who is an assistant professor of emergency medicine at Wake Forest University, has 12 years of army experience as a physician and special operations medic, and has augmented several special mission units as an emergency medicine physician. Prof. Masneri provided us with his best educated estimates for the expected triage time and expected time for stabilization, stressing that he was not aware of any studies on these times and that his responses were based on his experience and opinion only. (Stabilization here corresponds to service in our mathematical model.) It is also important to note that the estimates are based on the assumption that there is no longer fire exchange while triage and stabilization are performed. As a result of this consultation, we set the triage time in our study to 30 seconds and varied the stabilization time for the patients from four to eight minutes.

There are also scarcely any data that would allow highly reliable estimation of the death probability functions $f_1(t)$ and $f_2(t)$. The only available work to date that has attempted to make such estimation—partially relying on medical expert opinion—is that of Sacco et al. (2005, 2007) and Navin et al. (2009). In particular, Navin et al. (2009) provides on-site survival probability estimates for military-age victims with penetrating injuries (a type of trauma that is highly common in armed combat). We use these estimates, which are in fact provided at a more granular level than we need, to construct the death probability functions $f_1(t)$ and $f_2(t)$ we use in our study. For details on how we obtain these functions, see Section 7 of the online supplement. The estimates we obtain for $f_1(t)$ and $f_2(t)$ are plotted in Figure 3. Note that we can see from these plots that depending on the interval $[t_0, t_0 + \max_{i=0,1,2} N(\tau_i + u)]$, over which the least-square lines are fit for the heuristic policies we described in Section 6.2, the slopes of the fitted lines can be quite different. In particular, the ratio of the approximated $c\mu$ values for class 1 and class 2 patients, $(h_1/\tau_1)/(h_2/\tau_2)$, where h_i is the slope of the fitted line for $f_i(t)$, gets smaller as t_0, the time at which the response effort starts, increases. We will investigate how such a change in t_0, which essentially implies decreasing urgency of class 1 with respect to class 2, impacts the performances of our heuristic policies in Section 6.4.

To model the lifetimes, following Uzun Jacobson et al. (2012) and Hougaard (2012), we use Weibull distribution—i.e., we let $G_i(t) = 1 - e^{-(t/\beta_i)^{\theta_i}}$, where θ_i and β_i are shape and scale parameters, respectively, for class i patients ($i = 1, 2$). As in Uzun Jacobson et al. (2012), we use the time when $f_i(t)$ reaches some

Figure 3. Probability of Death Over Time for Penetrating Wound in a Battlefield



Note. The "immediate" category corresponds to *class 1* and the "delayed" category corresponds to *class 2* in our framework.

threshold η as the mean lifetime (starting from time zero) for class i patients and varied η from 0.90 to 0.99. In this paper, we only present the results for $\eta = 0.90$ (corresponding to mean life times of 60.96 and 423.61 minutes for classes 1 and 2, respectively) and $\eta = 0.95$ (corresponding to mean life times of 92.95 and 434.62 minutes for classes 1 and 2, respectively) since the results do not depend significantly on this choice. Following Uzun Jacobson et al. (2012), we let $\theta_1 = \theta_2 = 1.5$, and the scale parameters are computed using $\beta_i = m_i/\Gamma(1+1/\theta_i)$, $i = 1, 2$, where m_i denotes the mean lifetime for class i patients and $\Gamma(\cdot)$ is the incomplete gamma function.

Two parameters that are difficult to predict in advance are N, the total number of casualties, and α_1, the probability of a random casualty to be classified as immediate. In our study, we considered a range of values for both N and α_1 with N taking values from the set $\{5, 10, 15, 20, 25\}$ and α_1 taking values from the set $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. Soon after the event that triggers the mass-casualty situation takes place and response effort starts, the total number of casualties can be determined with a considerable degree of accuracy; however, α_1 may remain difficult to estimate. To investigate how our policies would perform in the case of such uncertainty, we also carried out a numerical study with a focus on the sensitivity of the performance of our policies to the reliability of the estimates for α_1.

6.4. Results of the Numerical Study

In this section, we compare the performances of the three heuristic policies we proposed in Section 6.2 and the two best benchmark policies analyzed in Section 5 (NT and TP1) with the performance of the optimal policy for (6). The performance measure of interest is the mortality rate, which is defined as the percentage of the total number of casualties who do not survive. For $N = 5, 10, 15, 20, 25$, Tables 1, 2, 3, 4, and 5, respectively, report the expected percentage increase that would be observed in the mortality rate by using one of the policies stated above instead of the optimal policy.

We can see that two of the policies we propose, DTP and STP-1, perform well in all of the scenarios with the percentage increase in the mortality rate (when compared with the optimal policy) mostly staying below 6%. There are only three scenarios in which the percentage difference exceeds 6% under DTP. More importantly, both DTP and STP-1, but particularly DTP, perform similarly or better than the benchmark heuristics NT and TP1. More specifically, they perform at least as good as or better than TP1 in all of the scenarios, and DTP performs similarly or better than NT in all of the scenarios except when N, the number of patients, is small. Out of the 150 different scenarios considered, NT outperforms DTP in only 12 scenarios, and these are all scenarios where there are few patients—i.e., $N = 5$. Note however that having few patients does not guarantee a good performance by NT. It appears that for NT to perform better than the other policies, not only the number of patients needs to be small but also α_1, the overall percentage of class 1 patients, needs to be high. On the other hand, when N is large (i.e., $N = 25$), one noteworthy but unsurprising observation is that the performances of our heuristic policies are similar to that of NT with identical performances observed when the mean service time τ is large. This is because when

Table 1. Percentage Increase in Mortality Rate by Using Heuristic Policies Over the Optimal Policy When $N = 5$

	$\eta = 0.90$ (%)					$\eta = 0.95$ (%)				
Heuristics	NT	TP1	STP-1	STP-2	DTP	NT	TP1	STP-1	STP-2	DTP
$\tau = 4$										
$\alpha_1 = 0.10$	2.61	1.73	1.73	1.73	1.73	1.96	1.39	1.39	1.39	1.39
$\alpha_1 = 0.30$	3.94	1.79	1.79	1.79	1.79	3.21	1.20	1.20	1.20	1.20
$\alpha_1 = 0.50$	2.51	2.30	2.30	2.30	2.30	2.06	1.51	1.51	1.51	1.51
$\alpha_1 = 0.70$	0.66	2.68	2.68	2.68	2.68	0.54	1.87	1.87	1.87	1.87
$\alpha_1 = 0.90$	0.00	3.61	3.78	3.78	3.78	0.00	2.87	3.06	3.06	3.06
$\tau = 6$										
$\alpha_1 = 0.10$	3.35	2.22	2.22	2.22	2.22	2.85	1.63	1.63	1.63	1.63
$\alpha_1 = 0.30$	5.32	2.87	2.87	2.87	2.87	4.75	1.96	1.96	1.96	1.96
$\alpha_1 = 0.50$	3.65	3.35	3.35	3.35	3.35	3.25	2.30	2.30	2.30	2.30
$\alpha_1 = 0.70$	1.38	3.32	3.32	3.32	3.32	1.22	2.40	2.40	2.40	2.40
$\alpha_1 = 0.90$	0.00	3.19	0.00	0.00	0.00	0.00	2.64	0.00	0.00	0.00
$\tau = 8$										
$\alpha_1 = 0.10$	3.13	2.87	2.87	2.87	2.87	3.17	2.05	2.05	2.05	2.05
$\alpha_1 = 0.30$	5.26	3.89	3.89	3.89	3.89	5.34	2.74	2.74	2.74	2.74
$\alpha_1 = 0.50$	3.80	4.37	4.37	4.37	4.37	3.74	3.09	3.09	3.09	3.09
$\alpha_1 = 0.70$	1.60	3.93	3.93	3.93	3.93	1.55	2.92	2.92	2.92	2.92
$\alpha_1 = 0.90$	0.00	2.96	2.09	2.09	2.09	0.00	2.49	1.71	1.71	1.71

Table 2. Percentage Increase in Mortality Rate by Using Heuristic Policies Over the Optimal Policy When $N = 10$

	$\eta = 0.90$ (%)					$\eta = 0.95$ (%)				
Heuristics	NT	TP1	STP-1	STP-2	DTP	NT	TP1	STP-1	STP-2	DTP
$\tau = 4$										
$\alpha_1 = 0.10$	4.91	9.44	2.18	9.44	3.63	3.76	9.45	1.32	9.45	2.33
$\alpha_1 = 0.30$	9.25	9.16	4.83	9.16	7.16	7.48	8.05	3.28	8.05	5.16
$\alpha_1 = 0.50$	7.35	8.68	4.04	8.68	4.93	6.06	7.47	2.66	7.47	3.36
$\alpha_1 = 0.70$	4.42	6.46	3.02	6.46	1.80	3.93	5.85	2.20	5.85	1.03
$\alpha_1 = 0.90$	2.79	4.17	2.79	2.79	2.79	3.07	4.28	3.07	3.07	3.07
$\tau = 6$										
$\alpha_1 = 0.10$	2.70	9.93	1.52	9.93	1.78	2.96	9.14	1.50	9.14	1.93
$\alpha_1 = 0.30$	7.22	11.42	4.54	11.42	5.73	7.13	10.06	4.09	10.06	5.46
$\alpha_1 = 0.50$	6.89	10.74	4.96	10.74	5.10	6.52	9.65	4.18	9.65	4.55
$\alpha_1 = 0.70$	4.61	7.52	4.10	7.52	2.64	4.31	6.93	3.39	6.93	2.14
$\alpha_1 = 0.90$	2.15	3.41	0.15	0.15	0.15	2.30	3.46	0.15	0.15	0.15
$\tau = 8$										
$\alpha_1 = 0.10$	0.93	11.18	1.13	11.18	0.17	1.78	9.40	1.37	9.40	0.95
$\alpha_1 = 0.30$	4.45	13.54	3.59	13.54	3.26	6.28	12.34	4.53	12.34	4.97
$\alpha_1 = 0.50$	5.24	12.21	4.86	12.21	3.81	6.41	11.60	5.16	11.60	4.83
$\alpha_1 = 0.70$	4.01	8.19	4.49	8.19	2.41	4.42	7.88	4.25	7.88	2.67
$\alpha_1 = 0.90$	1.89	3.28	1.80	2.52	0.17	2.01	3.24	1.67	2.46	0.14

there are many patients and it takes a long time to serve each patient, serving all of the patients is expected to continue for such a long time that the slopes of the linear approximations of the two survival probability functions end up being very close to each other, which in turn significantly reduces the potential benefits of triage and prioritization.

If we take a closer look at the comparison between the performances of DTP and STP-1, we can make a number of interesting observations. First, when the number of patients is small ($N = 5$), in almost all of the scenarios, all three policies we propose reduce to TP1, the policy of performing triage on all of the patients. When the number of patients is larger, however, the policies are no longer identical. In fact, the performance of TP1 gets significantly worse than DTP and STP-1 with the percentage difference with respect to the optimal policy being as high as 23% in some scenarios. When it comes to the comparison of DTP and STP-1, DTP appears to have a better performance overall, but

Table 3. Percentage Increase in Mortality Rate by Using Heuristic Policies Over the Optimal Policy When $N = 15$

	$\eta = 0.90$ (%)					$\eta = 0.95$ (%)				
Heuristics	NT	TP1	STP-1	STP-2	DTP	NT	TP1	STP-1	STP-2	DTP
$\tau = 4$										
$\alpha_1 = 0.10$	1.72	16.29	1.27	16.29	1.01	1.59	16.25	1.08	16.25	0.79
$\alpha_1 = 0.30$	7.83	15.48	5.79	15.48	6.61	6.84	15.03	4.82	15.03	5.46
$\alpha_1 = 0.50$	7.64	12.91	5.97	12.91	6.13	6.61	12.58	4.88	12.58	4.91
$\alpha_1 = 0.70$	4.77	8.29	4.22	8.29	3.06	4.15	8.15	3.47	8.15	2.25
$\alpha_1 = 0.90$	1.86	3.56	1.86	1.86	1.86	2.06	3.91	2.06	2.06	2.06
$\tau = 6$										
$\alpha_1 = 0.10$	0.55	16.60	1.27	16.60	0.02	1.07	15.00	1.42	15.00	0.48
$\alpha_1 = 0.30$	5.35	17.75	4.96	17.75	4.49	6.96	17.69	6.11	17.69	5.98
$\alpha_1 = 0.50$	6.76	14.74	6.39	14.74	5.66	7.62	15.07	6.85	15.07	6.40
$\alpha_1 = 0.70$	4.86	9.19	5.13	9.19	3.60	4.99	9.32	4.97	9.32	3.60
$\alpha_1 = 0.90$	1.65	3.17	1.65	1.65	1.65	1.67	3.26	1.67	1.67	1.67
$\tau = 8$										
$\alpha_1 = 0.10$	0.46	17.47	1.70	17.47	0.00	0.54	14.68	1.37	14.68	0.05
$\alpha_1 = 0.30$	2.43	18.76	3.43	18.76	1.72	5.95	19.24	6.17	19.24	5.16
$\alpha_1 = 0.50$	4.67	15.61	5.65	15.61	3.78	7.59	16.72	7.74	16.72	6.60
$\alpha_1 = 0.70$	4.09	9.70	5.16	9.70	3.07	5.31	10.23	5.78	10.23	4.19
$\alpha_1 = 0.90$	1.58	3.21	0.99	2.05	0.45	1.70	3.26	1.06	2.08	0.48

Table 4. Percentage Increase in Mortality Rate by Using Heuristic Policies Over the Optimal Policy When $N = 20$

	$\eta = 0.90$ (%)					$\eta = 0.95$ (%)				
Heuristics	NT	TP1	STP-1	STP-2	DTP	NT	TP1	STP-1	STP-2	DTP
$\tau = 4$										
$\alpha_1 = 0.10$	0.52	20.82	1.06	20.82	0.05	0.68	20.30	1.07	20.30	0.15
$\alpha_1 = 0.30$	6.01	17.95	5.25	17.95	5.19	6.08	18.50	5.27	18.50	5.15
$\alpha_1 = 0.50$	7.07	14.05	6.25	14.05	6.00	6.78	14.67	5.93	14.67	5.57
$\alpha_1 = 0.70$	4.57	8.48	4.34	8.48	3.33	4.22	8.77	3.93	8.77	2.83
$\alpha_1 = 0.90$	1.40	3.05	1.40	1.40	1.40	1.53	3.43	1.53	1.53	1.53
$\tau = 6$										
$\alpha_1 = 0.10$	0.37	19.92	1.25	19.92	0.00	0.40	17.78	1.01	17.78	0.00
$\alpha_1 = 0.30$	2.98	18.60	3.54	18.60	2.39	5.22	19.16	5.35	19.16	4.55
$\alpha_1 = 0.50$	5.56	15.04	5.89	15.04	4.78	7.46	16.36	7.39	16.36	6.58
$\alpha_1 = 0.70$	4.54	9.23	5.02	9.23	3.62	5.18	9.88	5.41	9.88	4.16
$\alpha_1 = 0.90$	1.40	2.90	1.40	1.40	1.40	1.40	3.00	1.40	1.40	1.40
$\tau = 8$										
$\alpha_1 = 0.10$	0.33	18.94	0.79	16.32	0.00	0.35	16.78	0.58	14.18	0.00
$\alpha_1 = 0.30$	1.07	19.26	2.23	19.26	0.58	4.05	19.51	4.61	19.51	3.51
$\alpha_1 = 0.50$	3.25	15.31	4.46	15.31	2.62	7.17	17.34	7.61	17.34	6.47
$\alpha_1 = 0.70$	3.63	9.58	4.62	9.58	2.89	5.48	10.64	5.93	10.64	4.66
$\alpha_1 = 0.90$	1.38	3.00	1.38	1.38	1.38	1.56	3.11	1.56	1.56	1.56

STP-1 still outperforms DTP in certain cases. Although there are some exceptions, generally, we observe the superior performance of STP-1 when the number of patients and the mean service times are small.

Overall, these numerical results suggest that when there are few patients in need of treatment, skipping triage altogether might be reasonable if the patients are more likely to be immediate than delayed. In all of the other cases (i.e., when the number of patients is not small or patients are not more likely to be immediate), we see significant benefits in performing triage. However, putting all of the patients through triage also does not work well. In fact, in the majority of the scenarios, it is much more preferable to skip triage altogether than to perform triage on all of the patients. Therefore, whether or not triage should be done on a patient should be determined carefully. Two of the policies we propose, which make this decision dynamically depending on the number of remaining patients, appear to work quite well. Of these two policies, DTP,

Table 5. Percentage Increase in Mortality Rate by Using Heuristic Policies Over the Optimal Policy When $N = 25$

	$\eta = 0.90$ (%)					$\eta = 0.95$ (%)				
Heuristics	NT	TP1	STP-1	STP-2	DTP	NT	TP1	STP-1	STP-2	DTP
$\tau = 4$										
$\alpha_1 = 0.10$	0.35	23.10	1.03	23.10	0.00	0.39	22.20	0.92	22.20	0.00
$\alpha_1 = 0.30$	3.94	17.71	3.79	17.71	3.33	4.50	18.59	4.27	18.59	3.81
$\alpha_1 = 0.50$	5.88	13.49	5.49	13.49	5.07	6.07	14.54	5.64	14.54	5.15
$\alpha_1 = 0.90$	1.14	2.68	1.14	1.14	1.14	4.00	8.59	3.90	8.19	2.91
$\alpha_1 = 0.70$	4.12	8.06	4.05	7.71	3.15	1.22	3.02	1.22	1.22	1.22
$\tau = 6$										
$\alpha_1 = 0.10$	0.29	20.60	0.65	18.20	0.00	0.31	18.88	0.50	16.44	0.00
$\alpha_1 = 0.30$	1.44	17.92	2.05	17.92	0.99	3.34	18.12	3.57	18.12	2.84
$\alpha_1 = 0.50$	4.01	13.89	4.50	13.89	3.41	6.41	15.58	6.51	15.58	5.74
$\alpha_1 = 0.70$	3.93	8.65	4.26	8.29	3.21	4.89	9.55	5.00	9.15	4.09
$\alpha_1 = 0.90$	1.21	2.63	1.21	1.21	1.21	1.22	2.74	1.22	1.22	1.22
$\tau = 8$										
$\alpha_1 = 0.10$	0.26	18.62	0.26	0.26	0.26	0.27	17.17	0.27	0.27	0.27
$\alpha_1 = 0.30$	0.92	19.07	0.92	0.92	0.92	2.55	18.17	2.55	2.55	2.55
$\alpha_1 = 0.50$	1.99	14.11	1.99	1.99	1.99	5.97	16.11	5.97	5.97	5.97
$\alpha_1 = 0.70$	2.93	8.89	2.93	2.93	2.93	5.12	10.16	5.12	5.12	5.12
$\alpha_1 = 0.90$	1.19	2.75	1.19	1.19	1.19	1.43	2.90	1.43	1.43	1.43

the policy that makes decisions based on both the number of unclassified patients and the number of patients classified as delayed, appears to work better overall. Policy STP-1, which is only described by a single threshold on the number of unclassified patients, also performs quite well. This good performance of STP-1 is important to highlight since simpler policies would have higher chances of being adopted in practice.

The numerical results reported so far were obtained under the assumption that the patients start being served or triaged right after the incident that caused the injuries (i.e., $t_0 = 0$). It is, however, possible that there could be some delays in starting the response effort due to various practical obstacles. This is important for comparison purposes because such a delay would imply that patients would have already "progressed" in their death probability curves, and the relevant portion of these curves, which are plotted in Figure 3, would not start at time zero but at some $t_0 > 0$. We next investigate how our results would change if t_0 were not zero. Specifically, we consider two cases: $t_0 = 10$ minutes and $t_0 = 30$ minutes, and we set $N = 10$. The results are provided in Table 6. We can observe from the table that our policies DTP and STP-1 continue to perform better than the simpler benchmark policies even when the response effort is delayed. It is worth noting, however, that the performance difference when compared with No-Triage policy is somewhat smaller. This is most likely a result of the fact when the starting time of the response effort is shifted, the slopes of the linear lines fitted to the death probability curves are closer to each other, which decreases the importance of classifying the patients. In fact, we observed that when t_0 is set to even a larger value that is greater than 60 minutes, the policies we propose, DTP, STP-1, and STP-2, all reduce to the No-Triage policy. In this case, the differences between the two classes are so small that it is not worth spending time to triage and prioritize patients. But perhaps more importantly, in practice, if the response effort starts that late, even "the best" policy could result in little benefit, as the probabilities of eventual death for all of the patients would have increased substantially.

Next we investigate the sensitivity of our results to the predicted value of α_1, the probability of a random casualty to be of the immediate class. It is reasonable to expect this probability to change from incident to incident, and it can be difficult to estimate. Therefore, it is important to investigate how badly our policies would perform if the policies were determined assuming a particular value of α_1 when in fact it is equal to something else. For this study, we repeated the scenarios we studied above. In these scenarios, the predicted value for α_1 was $\alpha_1^p \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. To investigate the sensitivity, we assumed that the true value of α_1 was not actually equal to the predicted value α_1^p but was a random variable uniformly distributed in the interval $(\alpha_1^p - \epsilon, \alpha_1^p + \epsilon)$, where $\epsilon = \min(\alpha_1^p, 1 - \alpha_1^p) \times 30\%$.

Table 7 reports the results for the percentage difference between the mortality rate under the optimal policy and that under each policy we investigate when $N = 10$. The 95% confidence intervals given in the table are based on 100 replications. We can observe that DTP and STP-1 collectively continue to perform well and better than the two benchmark policies. The mean percentage difference (with respect to the performance of

Table 6. Percentage Increase in Mortality Rate by Using Heuristic Policies Over the Optimal Policy When $N = 10$ and $\eta = 0.90$ and When the Response Operation Starts at t_0

	$t_0 = 10$ (%)					$t_0 = 30$ (%)				
Heuristics	NT	TP1	STP-1	STP-2	DTP	NT	TP1	STP-1	STP-2	DTP
$\tau = 4$										
$\alpha_1 = 0.10$	3.52	10.03	1.44	10.03	2.52	2.72	11.48	1.33	11.48	2.06
$\alpha_1 = 0.30$	7.17	9.71	4.09	9.71	5.76	6.44	11.91	4.81	11.91	5.82
$\alpha_1 = 0.50$	6.22	9.45	4.28	9.45	4.63	5.61	10.98	5.18	10.98	4.99
$\alpha_1 = 0.70$	3.94	7.06	3.55	7.06	2.23	3.42	7.81	4.20	7.13	2.81
$\alpha_1 = 0.90$	1.80	3.55	1.80	1.80	1.80	0.96	3.27	0.96	0.96	0.96
$\tau = 6$										
$\alpha_1 = 0.10$	1.92	11.06	1.14	11.06	1.14	1.78	13.06	1.36	13.06	1.19
$\alpha_1 = 0.30$	6.22	12.91	4.60	12.91	5.16	5.96	15.36	5.47	15.36	5.42
$\alpha_1 = 0.50$	6.07	11.74	5.30	11.74	4.87	5.79	13.42	6.24	13.42	5.27
$\alpha_1 = 0.70$	4.15	8.12	4.60	8.12	2.85	3.80	8.94	5.05	8.94	3.29
$\alpha_1 = 0.90$	1.65	3.34	1.65	1.65	1.65	1.18	3.41	1.18	1.18	1.18
$\tau = 8$										
$\alpha_1 = 0.10$	0.78	12.83	1.21	12.83	0.10	0.67	14.61	1.05	14.61	0.11
$\alpha_1 = 0.30$	4.06	15.41	4.07	15.41	3.18	4.05	17.82	4.98	17.82	3.56
$\alpha_1 = 0.50$	4.76	13.38	5.35	13.38	3.77	4.73	15.10	6.35	15.10	4.27
$\alpha_1 = 0.70$	3.63	8.81	4.96	8.81	2.56	3.45	9.66	5.34	9.66	2.99
$\alpha_1 = 0.90$	1.55	3.37	1.41	2.32	0.41	1.22	3.52	1.22	1.22	1.22

Table 7. Percentage Increase in Mortality Rate by Using Heuristic Policies Over the Optimal Policy When $N = 10$ and $\eta = 0.90$, 95% Confidence Interval Based on 100 Replications

	$\eta = 0.90$				
Heuristics	NT	TP1	STP-1	STP-2	DTP
$\tau = 4$					
$\alpha_1^p = 0.10$	4.45 ± 0.17	9.89 ± 0.07	1.93 ± 0.10	9.89 ± 0.07	3.22 ± 0.15
$\alpha_1^p = 0.30$	8.57 ± 0.02	9.45 ± 0.01	4.66 ± 0.01	9.45 ± 0.01	6.81 ± 0.02
$\alpha_1^p = 0.50$	6.91 ± 0.11	8.97 ± 0.06	4.15 ± 0.04	8.97 ± 0.06	4.93 ± 0.12
$\alpha_1^p = 0.70$	4.01 ± 0.04	6.80 ± 0.04	3.12 ± 0.01	6.80 ± 0.04	1.91 ± 0.04
$\alpha_1^p = 0.90$	2.21 ± 0.00	4.35 ± 0.00	2.21 ± 0.00	2.21 ± 0.00	2.21 ± 0.00
$\tau = 6$					
$\alpha_1^p = 0.10$	2.47 ± 0.11	10.21 ± 0.02	1.37 ± 0.06	10.21 ± 0.02	1.51 ± 0.11
$\alpha_1^p = 0.30$	6.79 ± 0.06	11.49 ± 0.03	4.40 ± 0.06	11.49 ± 0.03	5.45 ± 0.05
$\alpha_1^p = 0.50$	6.61 ± 0.06	10.87 ± 0.09	5.00 ± 0.01	10.87 ± 0.09	5.07 ± 0.07
$\alpha_1^p = 0.70$	4.39 ± 0.04	7.73 ± 0.06	4.24 ± 0.02	7.73 ± 0.06	2.73 ± 0.04
$\alpha_1^p = 0.90$	1.84 ± 0.00	3.60 ± 0.01	0.19 ± 0.00	0.19 ± 0.00	0.19 ± 0.00
$\tau = 8$					
$\alpha_1^p = 0.10$	0.97 ± 0.03	11.47 ± 0.01	1.19 ± 0.01	11.47 ± 0.01	0.14 ± 0.03
$\alpha_1^p = 0.30$	4.23 ± 0.08	13.49 ± 0.03	3.55 ± 0.08	13.49 ± 0.03	3.13 ± 0.07
$\alpha_1^p = 0.50$	5.03 ± 0.01	12.21 ± 0.12	4.84 ± 0.03	12.21 ± 0.12	3.76 ± 0.02
$\alpha_1^p = 0.70$	3.87 ± 0.03	8.30 ± 0.07	4.57 ± 0.02	8.30 ± 0.07	2.48 ± 0.03
$\alpha_1^p = 0.90$	1.72 ± 0.00	3.43 ± 0.01	1.90 ± 0.00	2.64 ± 0.01	0.23 ± 0.00

the optimal policy) under both policies is less than 6% in all of the scenarios except for one.

7. Conclusion

In emergency medicine, patient triage has largely been accepted as essential for a successful response effort. Especially in the chaotic scene that typically follows mass-casualty events, patient triage and prioritization helps in identifying those who would most benefit from emergency care and allocating resources accordingly. When the resources are extremely limited, however, to the extent that there is a single medic providing care on the scene, the wisdom of sticking with triage is questionable. Triage would certainly still help identify who should ideally be prioritized, but it is not clear whether delaying the actual treatment of the patients is worth that. This has been the central question of investigation in this paper, and our results strongly indicate that performing triage no matter what the conditions are could indeed make things worse.

Our findings suggest that when there are relatively few patients on the scene and patients are more likely to be immediate than delayed, it might be better to skip triage. Given that No-Triage policy performs relatively close to the optimal policy and better than the simple dynamic policies we propose, and that it is in general difficult to determine the "optimal" dynamic policy in practice, skipping triage may be advisable. Here, it is important to make it clear that the relatively quick triage at the very basic level with the sole goal of leaving *minor* and *expectant* patients out of consideration should continue but that the more lengthy process of further classifying patients as *immediate* or *delayed* could be skipped.

When the number of patients is not small, our results suggest that neither skipping triage completely nor performing triage on all of the patients works well. However, there are significant benefits to performing triage or skipping it depending on the system state (number of patients that are untriaged and triaged as low priority), and some of the relatively simple state-dependent policies we propose can help capture some of these benefits. These proposed policies are tested within a mathematical framework, which permits computation of the optimal policy and thereby a proper assessment of the performances of these policies. However, it is important to note that the policies are not custom designed for this specific framework and can be easily implemented in practice once the model parameters are properly estimated. Even though there are not much publicly available data on emergency response to mass-casualty events, the estimation should still be largely straightforward as most of the parameters such as mean service and triage times require data that are relatively easy to collect. The only exception to this is the survival probability functions, which are not only unknown but also difficult to estimate. Very few papers have dealt with the estimation of these functions (and remaining lifetime distributions) even though they are crucial not only in developing better quantitative methods for patient triage and prioritization but also for having better qualitative insights into the effects of delays on patient survival in the aftermath of mass-casualty events. Thus, estimation of survival probability functions along with remaining lifetime

probability distributions is a highly important avenue for future research.

One cause for concern when it comes to using the state-dependent policies we propose in practice could be whether it would be reasonable to expect that a paramedic on the scene would take the time to determine the policy to use. It could indeed be an unreasonable expectation depending on the nature of the event. However, the policy does not need to be determined on the scene after the event occurs. Such analysis can be done beforehand, and simpler guidelines based on our heuristic methods and mathematical analysis can be identified. During training, medics can be provided with these guidelines, which tell them what to do depending on the scene conditions such as the number of casualties they will need to take care of.

Finally, it is important to note that the main features of the decision problem we analyzed in this paper are relevant to many service systems in practice in addition to mass-casualty triage and prioritization. Some examples are search and rescue operations (Grissom et al. 2006, Genswein et al. 2008); internal maintenance and repair operations (Taghipour et al. 2011); prioritization of sales leads in marketing, particularly in business-to-business settings (Lichtenthal et al. 1989, Wilson 2003, D'Haen and den Poel 2013), where time is invested to assess the likelihood of existing leads to be successfully converted to actual sales; and intelligence (particularly human intelligence) collection management (Department of the Army 2006; Kaplan 2010, 2012; Ni et al. 2013), where agents make some initial investigation of existing ambiguous cues, which might possibly be pointing to potential terrorist activities, and prioritize them prior to more in-depth investigation. In fact, the decision problem at its core, that of balancing the time spent on acquiring more information with the time spent on acting on the available information, is not even unique to services. In our daily lives, we constantly prioritize our tasks by assessing the relative value of prioritizing one task over the other given the available information. In short, our mathematical analysis in this paper is more broadly relevant outside of the context of mass-casualty triage, and our results provide broader insights into making prioritization decisions in a large class of practical settings.

Acknowledgments

The authors thank David A. Masneri for generously sharing his expertise when setting the parameter values of the simulation model. The authors also thank the associate editor and the referees for their comments and suggestions, which significantly improved this paper.

References

Alizamir S, de Véricourt F, Sun P (2012) Diagnostic accuracy under congestion. *Management Sci.* 59(1):157–171.

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

Argon NT, Ziya S, Righter R (2008) Scheduling impatient jobs in a clearing system with insights on patient triage in mass casualty incidents. *Probab. Engrg. Inform. Sci.* 22(3):301–332.

Armony M, Gurvich I (2010) When promotions meet operations: Cross-selling and its effect on call center performance. *Manufacturing Service Oper. Management* 12(3):470–488.

Budhiraja A, Ghosh A, Liu X (2014) Scheduling control for Markov-modulated single-server multiclass queueing systems in heavy traffic. *Queueing Systems* 78(1):57–97.

Cox D, Smith W (1961) *Queues* (Methuen & Co., London).

Department of the Army (2006) *Human Intelligence Collector Operations* (Department of the Army).

D'Haen J, den Poel DV (2013) Model-supported business-to-business prospect prediction based on an iterative customer acquisition framework. *Indust. Marketing Management* 42(4):544–551.

Dobson G, Sainathan A (2011) On the impact of analyzing customer information and prioritizing in a service system. *Decision Support Systems* 51(4):875–883.

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

Genswein M, Thorvaldsdóttir S, Zweifel B (2008) Remote reverse triage in avalanche rescue. *Proc. Whistler 2008 Internat. Snow Sci. Workshop, Whistler, BC, Canada*, 63–72.

Gove S, Tamburlini G, Molyneux E, Whitesell P, Campbell H (1999) Development and technical basis of simplified guidelines for emergency triage assessment and treatment in developing countries. *Archives Disease Childhood* 81:473–477.

Grissom CK, Thomas F, James B (2006) Medical helicopters in wilderness search and rescue operations. *Air Medical J.* 25(1):18–25.

Güneş ED, Akşin OZ (2004) Value creation in service delivery: Relating market segmentation, incentives, and operational performance. *Manufacturing Service Oper. Management* 6(4):338–357.

Gupta AK, Smith KG, Shalley CE (2006) The interplay between exploration and exploitation. *Acad. Management J.* 49(4):693–706.

Gurvich I, Armony M, Maglaras C (2009) Cross-selling in a call center with a heterogeneous customer population. *Oper. Res.* 57(2): 299–313.

Harrison J (1975) Dynamic scheduling of a multiclass queue: Discount optimality. *Oper. Res.* 23(2):270–282.

Hougaard P (2012) *Analysis of Multivariate Survival Data* (Springer Science+Business Media, New York).

Kaplan EH (2010) Terror queues. *Oper. Res.* 58(4, Part 1):773–784.

Kaplan EH (2012) OR forum—Intelligence operations research: The 2010 Philip McCord Morse Lecture. *Oper. Res.* 60(6):1297–1309.

Klimov G (1974) Time-sharing service systems I. *Theory Probab. Appl.* 19(3):532–551.

Lerner EB, Schwartz RB, Coule PL, Weinstein ES, Cone DC, Hunt RC, Sasser SM, et al. (2008) Mass casualty triage: An evaluation of the data and development of a proposed national guideline. *Disaster Medicine Public Health Preparedness* 2(S1):S25–S34.

Lichtenthal JD, Sikri S, Folk K (1989) Teleprospecting: An approach for qualifying accounts. *Indust. Marketing Management* 18(1): 11–17.

Mabry R, McManus JG (2008) Prehospital advances in the management of severe penetrating trauma. *Critical Care Medicine* 36(7):S258–S266.

Mabry RL, Apodaca A, Penrod J, Orman JA, Gerhardt RT, Dorlac WC (2012) Impact of critical care—Trained flight paramedics on casualty survival during helicopter evacuation in the current war in Afghanistan. *J. Trauma Acute Care Surgery* 73(2):S32–S37.

Mandelbaum A, Stolyar AL (2004) Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Oper. Res.* 52(6):836–855.

March JG (1991) Exploration and exploitation in organizational learning. *Organ. Sci.* 2(1):71–87.

Mills A, Argon N, Ziya S (2013) Resource-based patient prioritization in mass-casualty incidents. *Manufacturing Service Oper. Management* 15(3):361–377.

Molyneux E, Ahmad S, Robertson A (2006) Improved triage and emergency care for children reduces inpatient mortality in a resource-constrained setting. *Bull. World Health Organ.* 84(4): 314–319.

Nain P (1989) Interchange arguments for classical scheduling problems in queues. *Systems Control Lett.* 12(2):177–184.

Navin DM, Sacco WJ, McGill G (2009) Application of a new resource-constrained triage method to military-age victims. *Military Medicine* 174(12):1247–1255.

Ni KS, Faissol D, Edmunds T, Wheeler R (2013) Exploitation of ambiguous cues to infer terrorist activity. *Decision Anal.* 10(1): 42–62.

Pinedo M (1983) Stochastic scheduling with release dates and due dates. *Oper. Res.* 31(3):559–572.

Posen HE, Levinthal DA (2012) Chasing a moving target: Exploitation and exploration in dynamic environments. *Management Sci.* 58(3):587–601.

Razzak JA, Kellermann AL (2002) Emergency medical care in developing countries: Is it worthwhile? *Bull. World Health Organ.* 80(11):900–905.

Sacco WJ, Navin DM, Fiedler KE, Waddell I, Robert K, Long WB, Buckman RF (2005) Precise formulation and evidence-based application of resource-constrained triage. *Acad. Emergency Medicine* 12(8):759–770.

Sacco WJ, Navin DM, Waddell RK, Fiedler KE, Long WB, Buckman RF Jr (2007) A new resource-constrained triage method applied to victims of penetrating injury. *J. Trauma Injury Infection Critical Care* 63(2):316–325.

Shumsky RA, Pinker EJ (2003) Gatekeepers and referrals in services. *Management Sci.* 49(7):839–856.

Smith W (1956) Various optimizers for single stage production. *Naval Res. Logist. Quart.* 3(1–2):59–66.

Stillman P, Strong P (2008) Pre-triage procedures in mobile rural health clinics in Ethiopia. *Rural Remote Health* 8:955.

Taghipour S, Banjevic D, Jardine A (2011) Prioritization of medical equipment for maintenance decisions. *J. Oper. Res. Soc.* 62(9): 1666–1687.

Ünlü A, Can MF, Yagci G, Ozerhan I, Asensio JA, Petrone P (2013) Tactical evacuation of casualties by military helicopters: Present and future aspects. *Panamerican J. Trauma Critical Care Emergency Surgery* 2(2):83–88.

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

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

Wang X, Debo LG, Scheller-Wolf A, Smith SF (2010) Design and analysis of diagnostic service centers. *Management Sci.* 56(11): 1873–1890.

Wilson RD (2003) Using online databases for developing prioritized sales leads. *J. Bus. Indust. Marketing* 18(4/5):388–402.

World Health Organization (2008) *Operations Manual for Delivery of HIV Prevention, Care and Treatment at Primary Health Centres in High-Prevalence, Resource-Constrained Settings* (WHO Press, Geneva).