M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 15, No. 3, Summer 2013, pp. 361–377
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.1120.0426

Resource-Based Patient Prioritization in Mass-Casualty Incidents

Alex F. Mills
Operations and Decision Technologies, Kelley School of Business, Indiana University, Bloomington, Indiana 47405, millsaf@indiana.edu

Nilay Tanık Argon, Serhan Ziya
Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599 {nilay@email.unc.edu, ziya@email.unc.edu}

The most widely used standard for mass-casualty triage, START, relies on a fixed-priority ordering among different classes of patients, and does not explicitly consider resource limitations or the changes in survival probabilities with respect to time. We construct a fluid model of patient triage in a mass-casualty incident that incorporates these factors and characterize its optimal policy. We use this characterization to obtain useful insights about the type of simple policies that have a good chance to perform well in practice, and we demonstrate how one could develop such a policy. Using a realistic simulation model and data from emergency medicine literature, we show that the policy we developed based on our fluid formulation outperforms START in all scenarios considered, sometimes substantially.

Key words: health operations; emergency response; triage
History: Received: September 23, 2011; accepted November 17, 2012. Published online in *Articles in Advance* January 28, 2013.

1. Introduction

Triage is the process of classifying patients according to their medical conditions and injury characteristics and then determining the order in which they will be treated or transported to a hospital. Triage is a common practice in emergency departments for daily emergencies, but it is particularly important in the aftermath of a mass-casualty incident (MCI). Mass-casualty incidents such as transportation accidents and terrorist bombings may create a sudden spike in demand for the emergency response resources within an area. As a result, even patients who are in critical condition may not have immediate access to these resources that are essential for their survival. In such an environment, it is thus crucial to allocate the limited resources to the patients in the most effective manner. Depending on the type of the incident, it might be necessary to make different types of resource allocation decisions at different stages of the response effort, such as on the site or at the hospital. Although these decision problems have many common features, this paper is particularly concerned with the on-site prioritization of patients for transportation to the hospital.

According to the current practice, resource allocation decisions at the site of an MCI are made in a very simple way: the triage class of a patient automatically determines the patient's priority. For example, consider the most widely adopted triage protocol in the United States, START, which stands for Simple Triage and Rapid Treatment (Lerner 2008). START classifies patients into four different classes. *Minor* patients are those who are capable of walking away from the scene; *delayed* patients are those with severe but not immediately life-threatening injuries; *immediate* patients are those with severe and immediately life-threatening injuries who can benefit from receiving treatment; and *expectant* patients are those whose injuries are so severe that they are expected to die even if substantial care is given. After the patients are classified, START gives the highest priority to patients in the immediate class, then to those in the delayed class. Once the incident site is cleared of patients in these time-critical classes, resources may be used for those in the minor and expectant classes.

There is a clear benefit from adopting a static, predetermined prioritization scheme that depends solely on patients' triage classes: it is simple and thus easy to implement. However, a number of recent papers in the emergency medicine literature have questioned the wisdom of this practice, primarily because it completely ignores resource limitations (Garner et al. 2001, Frykberg 2005, Jenkins et al. 2008). The main argument is that the priority levels of the patients should depend on the availability of resources relative to demand; in particular, in some cases it might be more sensible to give priority to patients who are in the delayed class as opposed to those in the

immediate class. For example, there may be so many immediate patients that there is insufficient time to get to the delayed patients within a time frame that would give them a good chance to survive. In that case, it might be better to give priority to the delayed patients first or to switch priority from the immediate patients to the delayed patients at some point during the response effort.

At an intuitive level, it may not be difficult to believe that taking resource limitations into account could improve the outcome of triage at an MCI. However, it is not clear exactly how this task should be carried out. Prior work has considered mainly two different approaches. (For a detailed review of these studies, see Argon et al. (2011).) One stream of work, which has mostly come out of the emergency medicine literature, aimed to develop policies that are computed in real time. In particular, Sacco et al. (2005, 2007) proposed the Sacco Triage Method (STM), which essentially solves a linear program immediately after the incident to determine the order in which the patients should be transported to the hospital. The STM requires collecting a significant amount of data after the incident, entering these data into a computer, and solving the associated linear program to determine the policy, all in a chaotic environment. Moreover, the examples provided by Sacco et al. (2005) show that the resulting priority policy can be quite complex, requiring switches from one priority class to another several times during the course of the response effort. Primarily because of these perceived impracticalities, the STM has not been well received by the medical community (Cone and MacMillan 2005).

The second approach uses mathematical modeling and analysis, mainly to generate insights and identify basic principles that would lead to practical resource-based prioritization policies. The most relevant work that has followed this approach is that of Argon et al. (2008) and Uzun Jacobson et al. (2012). In this line of work, the authors formulated a stochastic clearing model with a finite number of patients that are categorized into different criticality classes. In this model, each patient has a random lifetime whose probability distribution depends on the class of the patient. If a patient is not served before her lifetime, she dies; otherwise, she either definitely survives (Argon et al. 2008) or survives with a probability that depends on the class of the patient (Uzun Jacobson et al. 2012). In one respect, the way in which the cost of delay is captured by this formulation is very direct and realistic. However, an implicit assumption is that the survival probability of a patient stays at the same positive value as long as the patient is alive and drops to zero instantaneously when the patient dies. This assumed structure for the survival probability does not fit well with what is reported in the emergency medicine literature; see, e.g., Sacco et al. (2005).

In this paper, our main goal is not to propose a real-time solution method, but rather to carry out mathematical and numerical analysis to generate insights that would be useful in the design of effective yet simple prioritization policies. In that respect, our approach is closer to the second approach discussed above. Specifically, we aim to provide answers to some of the questions that the emergency response community may face in the process of developing resource-based prioritization policies. For example, is it possible to develop policies that are simple enough to be implemented in practice yet have substantial benefits over standard policies that do not consider resource limitations? If so, what are the main characteristics of these policies, and in what kind of mass-casualty incidents are they likely to be more beneficial?

Even though this paper is closer to the second line of work discussed above, our model is completely different from those of Argon et al. (2008) and Uzun Jacobson et al. (2012). More specifically, we develop a *fluid model* in which patients deteriorate over time according to a survival probability function. Thus, criticality is modeled through a diminishing reward function rather than through abandonments. We provide details of our model in §2. Using our fluid formulation, we first obtain an analytical characterization of the optimal policy, which helps us generate new insights into "good" patient prioritization policies (see §§3 and 4). We also use the optimal solution to our fluid formulation to design two simple resource-based prioritization policies that are compatible with START and other similar triage classifications, which have two time-critical patient classes (see §5). In §6, using a stochastic discrete-event simulator and data from the emergency medicine literature, we test the performance of our resource-based policies and gain further insights into the problem under more realistic conditions than those reflected by our fluid model. Finally, in §7, we explain how our policies can be adapted to potential changes on the field as a result of delays in the availability of patients, misclassifications during triage, and the practice of repeating triage in the midst of the response effort (which is usually called *retriage*), and we test the performance of these adaptive policies by another simulation study.

Before we proceed, it is important to note that the objective that we consider in this paper, i.e., the maximization of the expected number of survivors, is consistent with the widely accepted and practiced emergency response principle of *doing the greatest good for the greatest number* (Kennedy et al. 1996, Frykberg 2005). However, triage has always been a

somewhat contentious practice because it essentially entails favoring certain individuals over others. There is a long line of discussion and research on the ethical dimensions of triage and what its objective "should" be. For more on this issue, which is beyond the scope of this paper, we refer the reader to Winslow (1982), Baker and Strosberg (1992), and references therein.

2. Model Description

We consider a scenario where there are many injured patients who need to be transported to a hospital. In particular, we consider the case where ambulances or other transportation resources are limited in supply so that at least some of the patients will have to wait for some time before being transported. We assume that at time zero the patients have already been separated into N classes based on their injury characteristics and medical conditions, and moved to a single area of the site where they are given basic treatment and prepared for loading onto the ambulances. According to our formulation, there will be no new patient arrivals after time zero. Thus, our model is a better fit for incidents where a significant percentage of the patients are quickly accounted for and thus the response effort does not necessitate a time-consuming search and rescue activity. Nevertheless, in §7.2, we consider cases where some of the patients arrive with some delay by means of a simulation study. We denote the set of classes by $\mathcal{I} = \{0, 1, \ldots, N-1\}$, and the number of patients in class i by n_i, where $n_i > 0$. We also assume that all patients need to be transported to the same hospital via the same transportation mode (e.g., via ground transportation) so that the transportation time of a patient does not depend on the patient's class. For simplicity, we will frequently use the word "service" to refer to the process of transporting a patient to the hospital.

We approach this problem from the perspective of the emergency response coordinator, who decides the order in which patients should be transported, with the objective of maximizing the overall expected reward or gain from the system. To this end, we assume that each class i has an associated nonnegative reward function $f_i(t)$, which is the expected reward earned by the system if a class i patient is served at time t. To capture the fact that no patient would benefit from a delay in service, we assume that $f_i(t)$ is monotone nonincreasing in t for each $i \in \mathcal{I}$. For mathematical tractability, we further assume that the first-order derivative of $f_i(t)$ with respect to t exists for each $i \in \mathcal{I}$, and is denoted by $f_i'(t)$. The function $f_i(t)$ can be interpreted as the probability that a patient of class i ultimately survives if taken into service at time t. With this interpretation, maximizing the total expected reward is then equivalent to maximizing the expected number of survivors. Note that we do not explicitly model patients who die and leave the system. Instead, we implicitly model death through the survival probability. Because our objective is to maximize the expected number of survivors (more generally, the total expected reward), the optimal policy would be always such that patients with zero survival probability (dead patients) are the last patients to receive service. Thus, our formulation achieves some mathematical simplicity without sacrificing realism in any crucial way.

Our goal is to develop a model that captures the essential features of the patient prioritization problem but is simple enough to allow mathematical analysis and development of easy-to-implement policies that are expected to perform well in practice. Toward that end, we propose a fluid formulation where different classes of patients in the system correspond to different classes of fluid and service of those patients corresponds to a flow of the respective fluid out of the system. Without loss of generality, we assume that the service rate is one patient per unit time; therefore, when patients of only class i are flowing out of the system at time t, reward is earned at a rate of $f_i(t)$.

Define a set of decision functions $\mathbf{r}(t) \equiv \{r_i(t): [0, \infty) \to [0, 1], i \in \mathcal{I}\}$, where $r_i(t)$ is the rate at which we choose to serve class i patients or the fraction of the total service capacity allocated to class i patients at time $t \geq 0$. We restrict ourselves to decision functions that have finitely many discontinuities, which is needed to obtain solutions that switch priorities only finitely many times and hence are applicable in practice. We now state our optimization problem as follows:

$$\begin{aligned} \max_{\mathbf{r}(t),\, t\in[0,\,\infty)} \quad & \sum_{i=0}^{N-1} \int_0^\infty r_i(s) f_i(s)\, ds \\ \text{subject to} \quad & \sum_{i=0}^{N-1} r_i(t) \leq 1, \quad \forall\, t \in [0, \infty), \\ & \int_0^\infty r_i(t)\, dt = n_i, \quad \forall\, i \in \mathcal{I}. \end{aligned} \tag{P1}$$

We first note that, as one would expect, it is suboptimal to leave any of the available capacity unused as long as there is fluid in the system. (The proof is omitted because it immediately follows from the assumptions that the reward functions $f_i(t)$ are nonincreasing in t and there are no further arrivals.) The practical implication of this result is that in the rest of this paper we do not need to consider solutions that involve idling. Because the service rate is one patient per unit time, under nonidling policies the fluid will be cleared from the system, i.e., transportation of the patients will be complete, at time $T = \sum_{i\in\mathcal{I}} n_i$. Thus, we can restrict ourselves to the time interval $[0, T]$.

Our fluid formulation allows the total service capacity to be allocated to more than one patient class

at any particular point in time. The practical interpretation of such an allocation can be problematic because transportation vehicles cannot be allocated in a continuous manner. This would especially be difficult to deal with when there are few vehicles to allocate. However, the following proposition resolves this concern. (Proofs of Proposition 1 and all other propositions and theorems are provided in the online supplement, available at http://dx.doi.org/10.1287/msom.1120.0426.)

Proposition 1. *There exists an optimal solution to* (P1) *where only one class of patient is served at any given time.*

Proposition 1 implies that we can restrict the set of policies we consider to those which serve only one patient class at any point in time. For practice, this result suggests that at any point in time, there is only one highest-priority class, and all transportation resources available should be allocated to that class unless the number of such patients is less than the number of resources.

Proposition 1 is also useful technically because it allows us to consider a formulation that is equivalent to but easier to analyze than (P1). Define the set-valued decision variable $\mathbf{W} = \{W(i):\ i \in \mathcal{I}\}$, where $W(i)$ is the set of time points during which class i is served. Then, we can rewrite our optimization problem in the following way:

$$\begin{aligned} \max_{\mathbf{W}} \quad & \sum_{i=0}^{N-1} \int_{W(i)} f_i(t)\,dt \\ \text{subject to} \quad & \mu(W(i)) = n_i, \quad \forall\, i \in \mathcal{I}, \\ & \bigcup_{i=0}^{N-1} W(i) = [0, T], \\ & W(i) \cap W(j) = \varnothing, \quad \forall\, i \neq j, \end{aligned} \tag{P2}$$

where $\mu(W(i))$ is the total amount of time spent serving class i patients. In the rest of our analytical work, we will focus on this formulation. For both practical and technical reasons, we restrict ourselves to solutions $\mathbf{W}$ such that for each $i \in \mathcal{I}$, $W(i)$ can be partitioned into finitely many open intervals and possibly a set of zero measure.

3. A Simple Condition for Fixed-Priority Ordering

Many triage methods that are used in practice, such as START, assign a fixed priority to each class of patients. To be more precise, in a fixed-priority method, the triage class of each patient determines his or her priority level, which does not change with time throughout the response effort. Although several examples show that, in general, the optimal policy to (P2) is not a fixed-priority policy, i.e., the optimal policy is such that the priority ordering of the patient classes changes with time, because of the simplicity of fixed-priority policies, it is still important to investigate conditions under which the optimality of a such a policy is guaranteed.

It turns out that one condition that ensures an optimal fixed-priority relationship between two classes of patients is an ordering between the derivatives of their respective reward functions.

Proposition 2. *Suppose that there exist two classes, i and j, which have the property that*

$$f_j'(t) \leq f_i'(t), \quad \forall\, t \in [0, T]. \tag{1}$$

Given a feasible solution to (P2), *where some class i patients are served before some class j patients, there exists another solution where*

(i) *no class i patients are served before class j patients, and*

(ii) *the expected total reward obtained under the new solution is at least as large as the expected total reward obtained under the existing solution.*

Proposition 2 implies that if class j patients deteriorate at least as fast as class i patients over $[0, T]$, then there exists an optimal solution where class j has priority over class i at all times. To intuitively understand Proposition 2, it helps to think about the "opportunity cost" of delaying service to each class for a period of time. If the expected reward function of class j always decreases faster than that of class i, we will forego more expected reward by delaying the service of class j than we would by delaying the service of class i, for any arbitrary amount of time. The optimal policy is to delay the service of the class for which there is less to lose with time.

There are a few important points worth emphasizing regarding Proposition 2. First, the proposition does not assume an ordering between $f_i(\cdot)$ and $f_j(\cdot)$. This means that the deterioration rates, not the nominal values of the expected rewards (e.g., survival probabilities), determine dominance. For example, it is possible for class i patients to have lower survival probabilities than class j patients at all times and yet be assigned a lower priority than class j patients if their health conditions do not deteriorate as fast. Second, it is crucial to ensure that the ordering in (1) holds for all $t \in [0, T]$. One might expect that if $f_j'(t) \leq f_i'(t)\ \forall\, t \in [0, t_0]$ for some $t_0 < T$, then we could at least apply the result of the proposition for the time period $[0, t_0]$. Nonetheless, we can easily construct examples where this intuition does not hold. And third, it might seem at first that condition (1) does not depend on the scale of the mass-casualty incident, i.e., the number of patients of various classes, transportation capacity, etc. However, recall that T is the time the response

effort is over; hence, T increases with the number of patients and the response time per patient. This means that as the scale of the incident gets larger, T becomes larger and as a result it becomes increasingly less likely for (1) to hold.

Proposition 2 directly leads to the complete characterization of an optimal policy when all classes can be ordered according to condition (1):

Corollary 1. *Suppose that*

$$f'_{N-1}(t) \le f'_{N-2}(t) \le \cdots \le f'_0(t), \quad \forall\, t \in [0, T]. \tag{2}$$

Then there exists an optimal policy under which there is a fixed-priority ordering $(N-1, N-2, \ldots, 0)$ *among the patients so that for any* $i \in \{N-2, N-3, \ldots, 0\}$, *class* $i+1$ *patients have priority over class* i *patients*.

The policy prescribed by Corollary 1 is very simple and practical. Because the priority ordering is fixed, there is less room for mistakes during implementation. However, the optimality of the policy is only guaranteed under a relatively strong condition. Therefore, it is important to investigate what the optimal policy would be like when condition (2) does not hold.

We pursue this question in the following section, where we focus on the case with only two patient classes. There are two main reasons why we consider this particular scenario. First, characterization of the optimal policy—except under condition (2)—appears to be very difficult when there are more than two patient classes. However, more importantly, the two-class case fits perfectly well with the widely adopted triage classification used in START. As we described in §1, even though START puts patients into four classes, patients who are in the *expectant* class have almost no chance to survive, and those with *minor* injuries do not carry a risk of dying from their injuries. Therefore, the success of the response effort depends almost entirely on how priority decisions for patients in the *immediate* and *delayed* classes are handled. Hence, our analysis of the two-class case in the next section will help us explore how START can be expanded to include resource limitations using our formulation.

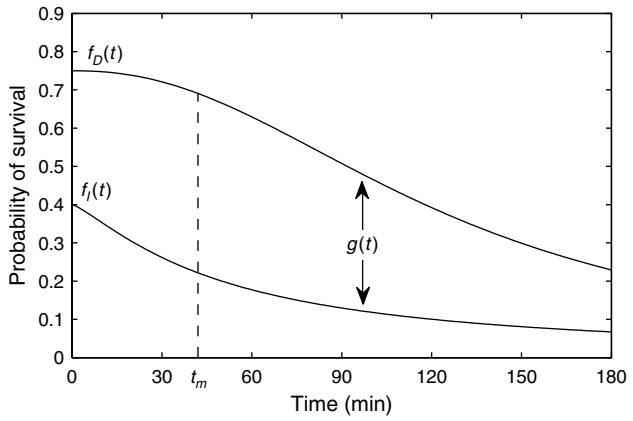
4. Priority Decisions for Two Classes of Patients

Suppose that there are only two patient classes, which we name class I (*immediate*) and class D (*delayed*). By letting $g(t) \equiv f_D(t) - f_I(t)$, we can rewrite the optimization problem (P2) as

$$\begin{aligned} \max_{W(D)} \quad & \int_{W(D)} g(t)\,dt + C \\ \text{subject to} \quad & \mu(W(D)) = n_D \\ & W(D) \subseteq [0, T], \end{aligned} \tag{P3}$$

where $C \equiv \int_0^T f_I(t)\,dt$ is a constant.

Figure 1 Example of Reward Functions Satisfying Assumption 1



From Proposition 2, we know that if patients of class I consistently deteriorate faster than patients of class D over $[0, T]$, then class I should have priority over class D at all times. Although this is a possibility in practice, a more likely scenario is one where patients with very critical conditions, who would be classified as *immediate* according to START, have rapidly diminishing survival probabilities, which approach zero fairly quickly, whereas the survival probability function for *delayed* patients would be relatively flat initially and then start to decline rapidly after some point. We formally state this scenario in Assumption 1 and present an example pair of reward functions, $f_I(t)$ and $f_D(t)$, satisfying Assumption 1 in Figure 1.

Assumption 1. *There exists* $t_m \in [0, T]$ *such that* $f'_I(t) < f'_D(t) < 0$ *for all* $t < t_m$, $f'_I(t_m) = f'_D(t_m)$, *and* $f'_D(t) < f'_I(t) < 0$ *for all* $t > t_m$. *Equivalently, the reward gap function* $g(t)$ *has a unique maximum at* $t_m \in [0, T]$, *is increasing for* $t < t_m$, *and is decreasing for* $t > t_m$.

We delay a more detailed discussion on the justification of Assumption 1 until §6.1, where we demonstrate that estimates for survival probability functions for immediate and delayed classes, which are based on data from Sacco et al. (2005), satisfy Assumption 1. The remainder of this section is devoted to characterizing the solution to (P3) under Assumption 1 and providing insights into the patient prioritization problem based on our analytical results.

4.1. Characterization of the Optimal Policy

In this section, we establish a useful structural result for the optimization problem (P3).

Proposition 3. *There exists an optimal policy in which* $W(D)$ *is a single interval*.

Proposition 3 implies that there is an optimal policy where once the service of delayed patients starts, it is never interrupted by the service of immediate

patients. Hence, the question of finding the optimal set of time points where delayed patients should be served reduces to the question of finding the time at which we should *begin* serving delayed patients. In other words, under Assumption 1, we can describe the optimal policy by a single time point, or threshold, and hence the optimization problem (P3) reduces to

$$\max_{t\in[0,\,n_I]} v(t), \tag{P4}$$

where $v(t) \equiv \int_t^{t+n_D} g(x)\,dx$. Let t^* denote a solution to problem (P4). Then, given t^*, an optimal policy can be described as follows: serve immediate patients over $(0, t^*]$, switch to delayed patients at t^* and serve all of them over the time interval $(t^*, t^*+n_D]$, and finally switch back to serving immediate patients and serve all the remaining immediate patients over the interval $(t^*+n_D, T]$. This general description encompasses the special case where the policy is a fixed-priority ordering policy like START. More specifically, if $t^* = n_I$, then all immediate patients are served before all delayed patients. On the other hand, if $t^* = 0$, then all delayed patients are served before all immediate patients. Threshold t^* can be interpreted as the time at which the "cost" of delaying service to delayed patients starts outweighing the "benefit" of providing service to immediate patients.

4.2. Determining the Optimal Threshold

We now investigate how one can obtain a solution to problem (P4). We first show that the solution t^* is unique, and it can be determined more easily by first solving a relaxation of (P4).

Proposition 4. (i) *There is a unique optimal solution, $\tilde{t}$, to the optimization problem* $\max_{t\in[0,\,\infty)} v(t)$, *and* $\tilde{t} \in [\max\{0, t_m - n_D\}, t_m]$.

(ii) $t^* = \min\{\tilde{t}, n_I\}$.

(iii) $t_m \in [t^*, t^*+n_D]$.

Part (i) of Proposition 4 partially characterizes $\tilde{t}$. Note that the difference between t^* and $\tilde{t}$ is that whereas t^* is restricted to be no greater than n_I, $\tilde{t}$ has no such restriction. Practically, $\tilde{t}$ is the time at which the service of delayed patients should start even if there are still immediate patients in need of service, which is very likely to be the case when there are many immediate patients initially. On the other hand, if all immediate patients are served before $\tilde{t}$, i.e., $n_I < \tilde{t}$, then service of delayed patients should start at n_I, because idling is suboptimal. Therefore, t^* is the minimum of $\tilde{t}$ and the time it would take to serve all immediate patients if they were given priority at all times. This is reflected in the relationship described in part (ii) of Proposition 4. Part (iii) of Proposition 4 states that the optimal service time interval for delayed patients must contain t_m. Thus, service of delayed patients should start late enough for the service interval to contain the time point at which deterioration of delayed patients becomes faster.

The following theorem provides a complete characterization of $\tilde{t}$, and thus a complete characterization of t^*, which subsequently leads to an algorithm for finding t^*.

Theorem 1. *Exactly one of the following three statements is true*:

(i) $g(0) > g(n_D)$, *in which case* $t^* = \tilde{t} = 0$.

(ii) $g(\tilde{t}) = g(\tilde{t}+n_D)$ *and* $g(n_I) \le g(T)$, *in which case* $t^* = n_I \le \tilde{t}$.

(iii) $g(\tilde{t}) = g(\tilde{t}+n_D)$ *and* $g(n_I) > g(T)$, *in which case* $t^* = \tilde{t} < n_I$.

Using Theorem 1, it is straightforward to show that the following algorithm determines the optimal threshold t^*:

1. If $g(0) > g(n_D)$, return $t^* = 0$.
2. Else, if $g(n_I) \le g(T)$, return $t^* = n_I$.
3. Else, return the solution of $g(t) = g(t+n_D)$.

Note that t^* is readily available if one of the two conditions in the first two steps holds. If neither holds, then one needs to determine the unique solution to $g(t) = g(t+n_D)$.

The first step in the algorithm checks whether $g(0) > g(n_D)$. This condition may hold when delayed patients deteriorate faster than immediate patients starting at either time zero or soon after, i.e., when t_m is close to or equal to zero. In this case, Theorem 1 indicates that $t^* = 0$, and hence delayed patients have priority at all times. This policy, which we call *Inverted START* (*InvSTART*), is the complete opposite of START because delayed patients have priority over immediate patients at all times. The same condition may also hold when n_D is very large, suggesting that InvSTART is optimal when there are sufficiently many delayed patients. If the condition in the first step of the algorithm is not satisfied, i.e., $g(0) \le g(n_D)$, then Theorem 1 states that $g(\tilde{t}) = g(\tilde{t}+n_D)$. In this case, t^* is equal to $\tilde{t}$ or n_I, depending on whether the inequality $g(n_I) > g(T)$ holds or not. If the algorithm stops in the second step, i.e., $t^* = n_I$, then the optimal policy is START because immediate patients are given priority over all delayed patients at all times. This case, where $g(n_I) \le g(T)$, may occur when either t_m is sufficiently large or the total number of patients is sufficiently small. Finally, if the algorithm stops in the third step, i.e., $t^* < n_I$, then the optimal policy is either InvSTART (when $t^* = 0$) or time dependent (i.e., priority will change at time $0 < t^* < n_I$). Under a time-dependent policy, the service of immediate patients is interrupted by the service of delayed patients during $(t^*, t^*+n_D]$.

4.3. Sensitivity of the Optimal Policy to the Number of Patients

From Theorem 1, it is clear that the optimal prioritization policy depends on the number of patients in

each class, because t^* is a function of both n_I and n_D. To better understand the relationship between the optimal policy and patient counts, we investigate how t^* changes with n_I and n_D.

By definition, $\tilde{t}$ does not depend on n_I. Hence, by part (ii) of Proposition 4, t^* increases with n_I for $n_I < \tilde{t}$ but does not change for $n_I \geq \tilde{t}$. This suggests that if there are few immediate patients and START is optimal (i.e., $t^* = n_I$), then increasing the number of immediate patients will not change the policy at first (except for the time when the service of delayed patients should start), but will eventually change the policy from START to a time-dependent one. However, when there are enough immediate patients for a time-dependent policy or InvSTART to be optimal, then having more immediate patients does not change the optimal policy.

We next present a proposition that describes how t^* depends on n_D.

Proposition 5. *Everything else remaining the same, t^* either decreases or stays the same as n_D increases, i.e., having more delayed patients can only decrease the time the service of delayed patients starts under the optimal policy.*

Proposition 5 and the discussion above yield the following conclusions:

(i) If the optimal policy is START, having more patients, regardless of their class, may push the optimal policy to be time dependent.

(ii) If the optimal policy is time dependent, having more immediate patients will not change the policy, whereas having more delayed patients will push the optimal policy toward InvSTART.

(iii) If the optimal policy is InvSTART, having more patients will not change the optimal policy.

5. A New Policy for Patient Triage: ReSTART

In this section, building on our analytical results from previous sections, we demonstrate how one could construct a new patient prioritization policy that takes into account resource limitations, yet is simple enough for practical implementation. More specifically, we carry the simple solution from §4 to practical settings where the fluid assumptions are obviously violated. We call the new policy *Resource-based START* (*ReSTART*) to indicate the fact that it builds on START, which is the most widely adopted triage method in United States. It is important to emphasize that ReSTART does not propose any new medical criteria to classify patients. ReSTART uses the START classes, but unlike START, it does not necessarily give priority to immediate patients at all times. Under ReSTART, delayed patients can get priority over immediate patients depending on the relative availability of the transportation vehicles with respect to the number of patients.

Now, let θ denote the expected transportation time for each patient, and let K denote the number of available transportation vehicles. Recall that in §4, we normalized the service rate to one, which is the same as assuming that $K/\theta = 1$. Incorporating generality in service rates, i.e., allowing a general number of vehicles and general transportation times, simply requires scaling of the number of patients by θ/K. The description below is based on the algorithm given in §4.2.

Resource-based START:

1. Classify patients according to the START classes.

2. Determine the number of patients classified as immediate (n_I) and the number of patients classified as delayed (n_D). Determine θ, the expected round-trip travel time for each transportation vehicle, and K, the number of vehicles that can be used for transporting patients to the hospital.

3. Determine priorities among the immediate and delayed patients as follows:

(i) If $g(0) > g(n_D\theta/K)$, transport all delayed patients first, followed by all immediate patients.

(ii) If $g(n_I\theta/K) \leq g((n_I + n_D)\theta/K)$, transport all immediate patients first, followed by all delayed patients.

(iii) Otherwise, determine t^* such that $g(t^*) = g(t^* + n_D\theta/K)$. Transport immediate patients until time t^* or until there are no more remaining immediate patients. Then, start transporting delayed patients and continue until there are none remaining. Finally, continue with the transportation of any remaining immediate patients.

Given the reward functions $f_I(\cdot)$ and $f_D(\cdot)$, implementation of ReSTART is straightforward to a large extent. Although in steps 3(i) and 3(ii), one only needs to check whether the given inequalities hold, in step 3(iii) a solution to $g(t^*) = g(t^* + n_D\theta/K)$ must be found. Because the right-hand side of the equation depends on n_D, θ, and K, t^* can be computed only after the incident occurs; in other words, it cannot be computed "off-line." The computation of t^* can be done very quickly using a line-search algorithm, because $g(\cdot)$ is a unimodal function. Nevertheless, this cannot necessarily be done by hand, which might be a cause for resistance to any potential implementation of ReSTART. Therefore, we simplify the policy further by proposing an approximation for t^* based on our analytical results. To distinguish this approximate version of ReSTART from the exact version described above, we call it *Quick-ReSTART* (*Q-ReSTART*). In particular, we propose two different versions of Quick-ReSTART, which we call *QuickDynamic-ReSTART* (*QD-ReSTART*) and *QuickStatic-ReSTART* (*QS-ReSTART*). In the following, we describe these two policies.

5.1. QuickDynamic-ReSTART

QD-ReSTART is essentially the same as ReSTART except that it does not use the exact value of t^*, but rather an approximation for t^*. From Proposition 4, with the proper scaling for the expected travel time and the number of ambulances, we know that $\tilde{t} \in [t_m - n_D\theta/K, t_m]$. Therefore, even if we cannot locate $\tilde{t}$ exactly, it would be reasonable to expect that approximating $\tilde{t}$ with a choice from this interval could lead to a policy that performs well. Now, we know that there exists $\tilde{\phi} \in [0,1]$ such that $\tilde{t} = t_m - \tilde{\phi} n_D\theta/K$. Instead of determining $\tilde{\phi}$ exactly, we approximate it by some $\phi \in [0,1]$. Then, we use $\tau = t_m - \phi n_D\theta/K$ instead of $\tilde{t}$ and set $t^* = \min\{n_I, \tau\}$ by part(ii) of Proposition 4. We call the policy that uses this approximation *QD-ReSTART*(ϕ). Thus, QD-ReSTART is in fact a family of policies, with each policy being uniquely described by a value of the parameter $\phi \in [0,1]$. Below is a formal description of QD-ReSTART.

QD-ReSTART(ϕ):

1. Same as ReSTART.
2. Same as ReSTART.
3. Compute $\tau = t_m - \phi n_D\theta/K$ and prioritize the immediate and delayed patients as follows:

(i) If $\tau \le 0$, transport all delayed patients first, followed by all immediate patients.

(ii) If $\tau \ge n_I\theta/K$, transport all immediate patients first, followed by all delayed patients.

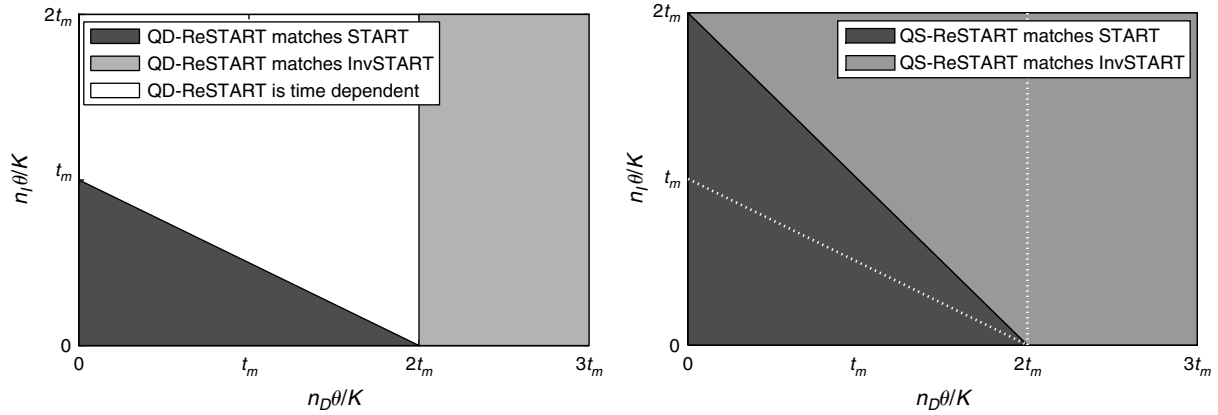
(iii) If $0 < \tau < n_I\theta/K$, transport immediate patients until time τ or until there are no more remaining immediate patients. Then, start transporting delayed patients (if there are any) and continue until there are no remaining delayed patients. Finally, continue with the transportation of any remaining immediate patients.

As a result of using τ in place of $\tilde{t}$, the inequalities in steps 3(i) and 3(ii) of ReSTART simplify to conditions that are easier to check and have insightful interpretations. Here, τ can be interpreted as a measure of the availability of resources relative to the size of the event. It is larger when there are fewer delayed patients, when transportation times are shorter, and/or when more vehicles are available for transportation. Thus, lower values of τ indicate more serious resource limitations.

Although ReSTART is the optimal policy for our fluid model under Assumption 1 and is likely to perform better than QD-ReSTART even under realistic conditions where the fluid assumption is relaxed, QD-ReSTART is simpler and more practical. Like ReSTART, it requires estimates for the expected travel time, number of ambulances, and number of immediate and delayed patients, which should not be difficult to determine, and which are likely to be the minimal set of requirements for any policy that takes resource limitations into account. However, unlike ReSTART, QD-ReSTART requires only an arithmetic calculation at the time of implementation. It simply uses t_m and ϕ, which can be obtained off-line before the incident based on estimates for the reward functions. Furthermore, τ depends on the function $g(\cdot)$ only through its maximizer t_m, meaning that the only estimation required is for the time at which the deterioration rate of the delayed patients exceeds that of the immediate patients.

To understand how QD-ReSTART works, it is useful to examine the leftmost plot in Figure 2, which depicts the structure of QD-ReSTART(0.5). In this plot, the horizontal axis is for $n_D\theta/K$, which is the expected time it would take to transport all delayed patients, and the vertical axis is for $n_I\theta/K$, which is the expected time it would take to transport all immediate patients using the full transportation capacity. It is immediate from the figure that for given values of expected transportation time and number of available vehicles, the priority ordering according to QD-ReSTART depends on the initial number of immediate and delayed patients. When there are few patients of both classes (in the triangular region at the bottom left), QD-ReSTART reduces to START, giving priority

Figure 2 Visualizations of QD-ReSTART(0.5) and QS-ReSTART(0.5)



to immediate patients until they are all transported. When there are sufficiently many delayed patients (far right in the figure), regardless of the number of immediate patients, the priority is reversed: QD-ReSTART reduces to InvSTART, and transportation of immediate patients starts after all delayed patients are transported. On the other hand, when there are sufficiently many patients but the number of delayed patients is below a certain level, priority ordering changes with time; immediate patients have priority initially, but at some specified time, priority moves to delayed patients even if there are still immediate patients waiting. Those remaining immediate patients wait until all the delayed patients are transported. Note that this structure for QD-ReSTART is consistent with the behavior of the optimal solution to the fluid model (see §4.3).

One remaining question is how to set the value for ϕ. More empirical work is needed to make a more confident decision about this, but we demonstrate in §6 that for the survival probability data that we use in our simulation experiments, setting $\phi = 0.5$ provides a good performance for the QD-ReSTART policy. For more on the justification of the use of $\phi = 0.5$, see §6.1.

5.2. QuickStatic-ReSTART

QD-ReSTART is a dynamic policy in the sense that the class that has the higher priority can change as time passes during the response effort. Although this priority switch can only happen once, one might still want to use even a simpler policy that fixes the priority levels at the beginning and does not change them later on. More precisely, one can choose either START or InvSTART given the conditions at time zero and use it until all patients are transported. We propose such a policy, which we call QS-ReSTART(ϕ), based on our analytical characterization of ReSTART, more specifically, QD-ReSTART(ϕ).

We can observe from Figure 2 that QD-ReSTART is a time-dependent policy in only one of the three regions. To develop QS-ReSTART(ϕ), we simply divide this region into two with a line that passes through the points $(t_m/\phi, 0)$ and $(0, t_m/\phi)$, merge the left part with the already existing "START" region, and merge the right part with the already existing "InvSTART" region, thereby eliminating the time-dependent policy region completely (see the rightmost plot in Figure 2, where $\phi = 0.5$). The policy can then be described simply as follows: use START if $n_I + n_D \le Kt_m/(\phi\theta)$, and use InvSTART otherwise. Note that one nice feature of this policy is that whether START or InvSTART is chosen depends on the total number of patients $n_I + n_D$, not on n_I and n_D individually. This means that once the total number of patients is determined, the policy can be determined even before triage is over. Furthermore, the policy is robust to classification errors between immediate and delayed classes because such errors would not change the total number of critical patients.

6. Simulation Study

In this section, we carry out a simulation study to investigate how QD-ReSTART and QS-ReSTART perform compared with START and InvSTART under conditions that are more realistic than those of the fluid model. Specifically, we study a simulation model where patients are discrete entities, ambulances are discrete resources, and transportation times are stochastic. In §6.1, we provide details on our experimental setup. Our results on the comparison of QD-ReSTART and QS-ReSTART with START and InvSTART are provided in §6.2. In §6.3, we present a sensitivity analysis with respect to the reward functions that are used in our simulation study.

6.1. Experimental Setup

We consider a mass-casualty incident that takes place at a single location and results in a number of patients who need to be transported to a hospital. We assume that the patients have already been classified and they are ready to be transported. The initial arrival of the ambulances to the site follows a Poisson process. For each ambulance, we assume that the round-trip travel time between the incident location and the hospital has lognormal distribution with a mean of 30 minutes and a standard deviation of 12 minutes. This choice is based on an empirical study by Ingolfsson et al. (2008) that reports that a lognormal distribution with standard deviation that is equal to 40% of the mean is a good fit for ambulance travel times.

We use the *critical mortality rate*, i.e., the percentage of critical patients who die, as the performance measure of interest. Frykberg (2005) stated that critical mortality is the best measure of performance for mass-casualty triage because it takes into account only those patients whose conditions are serious enough to require timely treatment and who also have a nonnegligible chance of survival. Note that because these are the only patients for whom a priority policy can make a difference, minimizing the critical mortality rate is equivalent to maximizing the number of survivors.

In our numerical experiments, we assume that patients are categorized according to START guidelines, because START is the most widely accepted classification method in the United States. Recall that there are four classes of patients according to START: expectant (E), immediate (I), delayed (D), and minor (M). Patients who fall into the immediate and delayed classes are considered critical patients. To use critical mortality rate as our performance measure, we

need estimates for the survival probabilities of these critical patients as functions of time. To the best of our knowledge, the only work that attempted to estimate survival probability functions is that of Sacco et al. (2005, 2007), where the estimates are for a given initial RPM (Respiration, Pulse, and Motor response) score of a patient.

To obtain estimates of survival probability functions for critical START classes (i.e., immediate and delayed classes), we utilized the RPM score-based estimates by Sacco et al. (2007) and also consulted James E. Winslow, M.D., an associate professor of emergency medicine at Wake Forest University (Winslow 2010). Our analysis revealed that the following three-parameter function is a good model for the reward functions $f_I(t)$ and $f_D(t)$:

$$f_i(t) = \frac{\beta_{0,i}}{(t/\beta_{1,i})^{\beta_{2,i}} + 1}, \quad \text{for } i \in \{I, D\}, \tag{3}$$

where $\beta_{j,i} > 0$ for $j = 0, 1, 2$ and $i \in \{I, D\}$. Note that this function is a scaled version of the log-logistic distribution, which is commonly used in survival analysis (Cox and Oakes 1984). Furthermore, among such functions, this function provided one of the best fits to the empirical data that originated from Sacco et al. (2007). The online supplement discusses the details of this curve fitting study and how we estimated the parameters of model (3) using the data from Sacco et al. (2007) and Winslow (2010).

To our knowledge, there is no existing research on the survival probability functions or patient distributions for START. Therefore, rather than construct just one scenario that we claim to be correct, in our numerical study, we consider five possible scenarios, which mainly differ according to how "pessimistic" they are regarding the size of the mass-casualty incident and the urgency of patients. Each scenario is based on a different underlying distribution of the severity of the patients, and so each scenario has a different set of survival probability functions and a different patient distribution (additional details are given in the online supplement). More specifically, in Scenario 1, the survival probabilities are low, and most of the patients are in immediate or expectant categories; in Scenario 5, survival probabilities are much larger, and there are not many immediate or expectant patients; and Scenarios 2–4 are in the middle in terms of the severity of the event. Figure 3 explicitly shows the differences among these five scenarios. In the first column, we provide the empirical data on the survival probabilities, the fitted functions given by (3), and also t_m values. We observe that as we move from Scenario 1 to Scenario 5, survival probabilities at any given time and t_m increase. In the second column, we plot the difference between the two fitted survival probability functions, i.e., $g(t)$, which shows that Assumption 1 holds over the time interval of interest. Finally, the third column shows the probability distributions for the START classes. We can observe that although immediate patients constitute the highest percentage in most of the scenarios, their frequency decreases as we move from Scenario 1 to Scenario 5, indicating a decreasing level of criticality in the patient population.

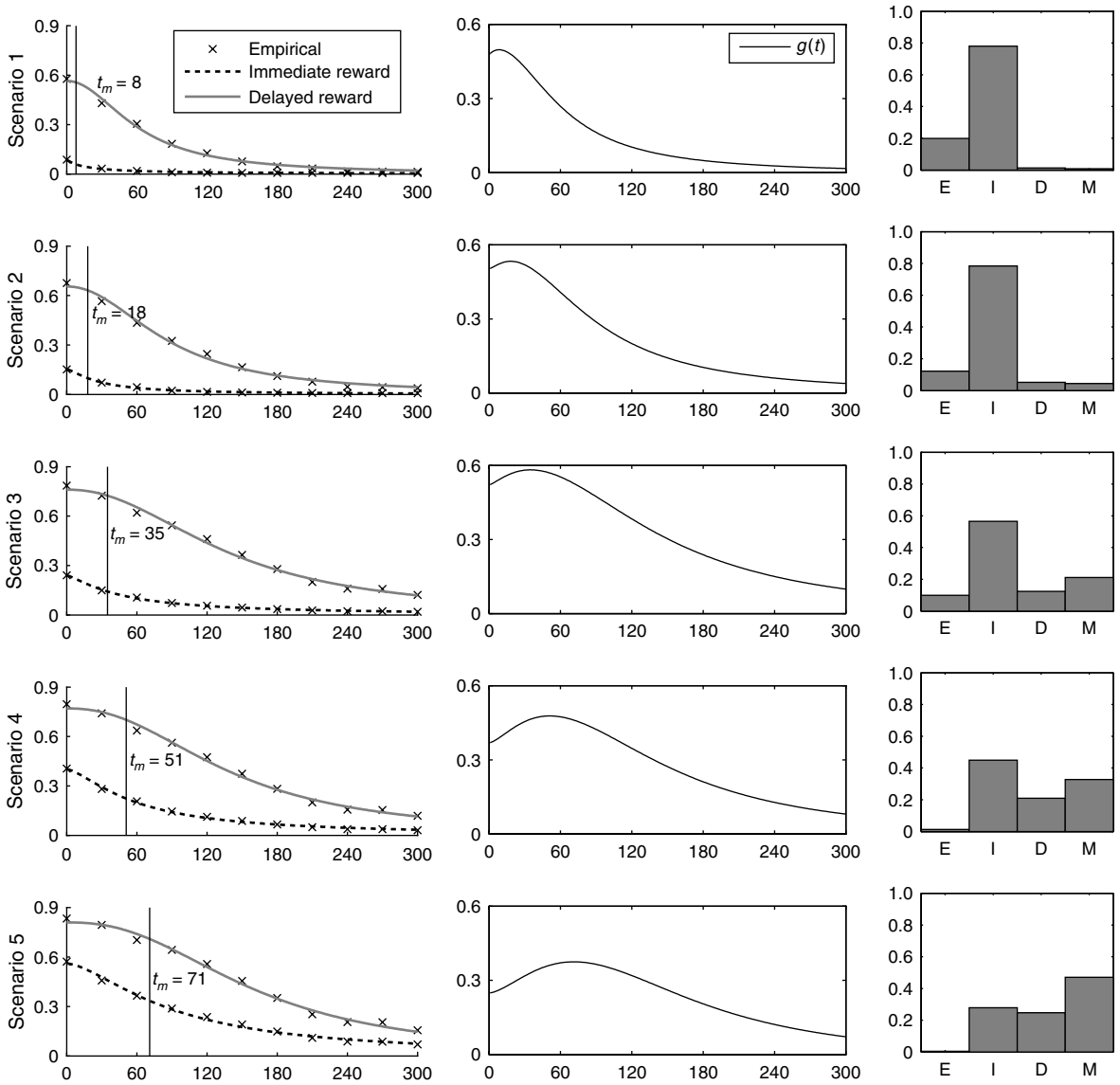
For each scenario, we generated 500 instances of 50 patients each, where each patient's START class was randomly determined using the START class probability distribution associated with that scenario. We used three different values for K, the number of ambulances available for transportation: 5, 10, and 15. The rate of initial arrival for ambulances to the event location was chosen as 10, 20, and 30 per hour when there are 5, 10, and 15 ambulances available for transportation, respectively. By keeping the number of patients and expected travel times constant but varying the number of resources, we were able to examine scenarios that ranged from resource-scarce to resource-abundant.

Within each scenario and resource level, we determined the performances of four policies, namely, START, InvSTART, QD-ReSTART, and QS-ReSTART, for each of the 500 randomly generated instances, using 200 replications for each instance. (To perform these simulation runs, we used code written in Matlab.) To implement QD-ReSTART and QS-ReSTART, we needed to set an appropriate value for ϕ. From Figure 3, we observed that $g(t)$ is almost symmetric around t_m over $[0, 2t_m]$, which means that setting ϕ to 0.5 will yield $\tau \approx \tilde{t}$, i.e., QD-ReSTART(ϕ) will be approximately identical to ReSTART. We also calculated the value of $\tilde{\phi}$ that would make $\tau = \tilde{t}$ for each instance generated where t^* is nonzero. We then constructed 95% confidence intervals on the mean values of such $\tilde{\phi}$'s for all five scenarios and three levels of resource availability. The means of these 15 confidence intervals ranged between 0.42 and 0.51, and the half-lengths of these intervals were all less than 0.01. Based on these observations, we set ϕ to 0.5 in all our simulation experiments that involve QD-ReSTART and QS-ReSTART.

6.2. Comparison of QD-ReSTART and QS-ReSTART with START and InvSTART

We first compare QD-ReSTART with START and InvSTART in terms of the critical mortality rate in Table 1. In this table, the sixth and eleventh columns report the number of instances (out of 500) in which the mean performance difference is statistically significant at the 0.05 level. The numbers in the second through fifth and seventh through tenth columns provide a statistical summary of the mean reduction in critical mortality obtained by using QD-ReSTART as opposed to using START

Figure 3 Survival Probability (Reward) Functions, $g(t)$, and START Class Distributions for the Five Scenarios



and InvSTART, respectively, based on 500 simulated instances: the minimum improvement (min), the maximum improvement (max), the mean, and the 95% half-width for the mean (HW).

From Table 1, we observe that QD-ReSTART performs at least as well as START in all instances while performing better in many. The magnitude of the improvement depends on the scenario considered. We cannot claim any one scenario as being more realistic than the others; however, we can still make a few insightful observations. In Table 1, going from Scenario 1 to Scenario 5, by which the survival probability estimates become more "optimistic," performance improvement with QD-ReSTART first increases and then decreases. The most significant improvement is in Scenarios 2–4. In Scenario 5 with a large number of ambulances, which is the most optimistic case considered in this study, the performance improvement is the smallest. This is because under Scenario 5, delaying transportation of the delayed patients affects them the least, so one can "afford" to use START by transporting all immediate patients first before moving on to the delayed patients. In Scenario 1, on the other hand, survival probabilities for both types of patients are so low that there is not much room for reduction in critical mortality. As a result, the difference between the performances of *any* two priority policies cannot be large. Nevertheless, even under such a pessimistic scenario, the mean improvement by QD-ReSTART is statistically significant regardless of the number of ambulances.

The number of available ambulances also has a clear effect on the performance improvement achieved by QD-ReSTART. Under all scenarios, the

Table 1 Mean Reduction in the Critical Mortality Rate Obtained Using QD-ReSTART Instead of START or InvSTART

	QD-ReSTART vs. START					QD-ReSTART vs. InvSTART				
K	Min (%)	Max (%)	Mean (%)	HW (%)	No. of sig. instances	Min (%)	Max (%)	Mean (%)	HW (%)	No. of sig. instances
Scenario 1										
5	0.0	5.8	0.7	0.1	223	0.0	0.0	0.0	0.0	0
10	0.0	4.8	0.5	0.1	223	0.0	0.0	0.0	0.0	0
15	0.0	3.5	0.4	0.0	223	0.0	0.0	0.0	0.0	0
Scenario 2										
5	0.0	8.8	2.8	0.2	465	0.0	0.1	0.0	0.0	0
10	0.0	6.5	2.0	0.1	465	0.0	0.1	0.1	0.0	254
15	0.0	4.1	1.2	0.1	465	0.0	0.2	0.1	0.0	333
Scenario 3										
5	0.0	11.4	6.2	0.2	499	0.0	0.4	0.2	0.0	392
10	0.0	4.5	2.1	0.1	499	0.0	0.8	0.6	0.0	499
15	0.0	1.5	0.6	0.0	486	0.0	1.1	0.7	0.0	498
Scenario 4										
5	2.3	10.3	6.8	0.1	500	0.0	1.3	0.6	0.0	422
10	0.1	3.1	1.1	0.0	484	1.1	2.6	1.8	0.0	500
15	0.0	0.5	0.1	0.0	15	1.2	2.9	2.3	0.0	500
Scenario 5										
5	0.0	5.7	2.2	0.1	478	0.0	3.2	1.5	0.0	494
10	0.0	0.3	0.0	0.0	0	0.9	3.7	3.3	0.0	500
15	0.0	0.0	0.0	0.0	0	0.8	3.6	2.8	0.0	500

improvement with QD-ReSTART is larger when there are fewer ambulances. When there are many ambulances available, even when all immediate patients have priority over all delayed patients, the transportation of delayed patients will not be delayed very long. However, when resources are scarce, by switching the priority to delayed patients after a certain period of time, QD-ReSTART saves more of these delayed patients, who would otherwise have a lower chance of survival when they are transported.

From Table 1, we also observe that the performance of QD-ReSTART is at least as good as that of InvSTART in all instances considered. Perhaps not surprisingly, in contrast with START, InvSTART does better in pessimistic scenarios, with a performance matching that of QD-ReSTART. In the more optimistic scenarios, InvSTART performs poorly, especially when there are many ambulances. This is because InvSTART transports delayed patients first at the expense of delaying immediate patients, even though delayed patients can actually "afford" to wait longer.

Finally, we compare QS-ReSTART with START, InvSTART, and QD-ReSTART in terms of the critical mortality rate in Table 2. We can observe that the mean improvements over START and InvSTART using QS-ReSTART are smaller than those using QD-ReSTART, but not drastically. This same observation is reflected in the negative values in the direct comparison with QD-ReSTART: QS-ReSTART performs worse than QD-ReSTART, but only by a relatively small amount. This suggests that QS-ReSTART would be a reasonable alternative to QD-ReSTART if time-dependent priority levels turn out to be difficult to implement in practice. Note that under each scenario, for any given number of ambulances, at least one of START or InvSTART performs very similarly to QS-ReSTART. This is expected because QS-ReSTART essentially chooses one of the two.

6.3. Sensitivity Analysis on the Reward Functions

To evaluate whether similar performance improvement could be achieved even without precise knowledge of the reward functions, we conducted a sensitivity analysis for the β parameters in the fitted reward functions given by (3). We repeated the simulations on the same instances used in the study presented in §6.2. However, this time, for each instance, we randomly perturbed the time-zero probability β_0, the scale parameter β_1, and the shape parameter β_2, for both reward functions that are used as inputs to the simulation runs, but we did not perturb the reward functions while determining the operating parameters of the QD-ReSTART(0.5) and QS-ReSTART(0.5) policies (i.e., t_m). This way, we were able to test the performance of QD-ReSTART and QS-ReSTART policies when the estimates for reward functions were not accurate.

We considered two experimental settings for the perturbations. In Setting 1, each β parameter was equally likely to decrease 10%, decrease 5%, stay the same, increase 5%, or increase 10%, whereas in Setting 2, each β parameter was equally likely to decrease 20%, decrease 10%, stay the same, increase

Table 2 Mean Reduction in the Critical Mortality Rate Obtained Using QS-ReSTART as Opposed to START, InvSTART, or QD-ReSTART (Negative Numbers Indicate an Increase in Critical Mortality)

	QS-ReSTART vs. START			QS-ReSTART vs. InvSTART			QS-ReSTART vs. QD-ReSTART		
K	Mean (%)	HW (%)	No. of sig. instances	Mean (%)	HW (%)	No. of sig. instances	Mean (%)	HW (%)	No. of sig. instances
Scenario 1									
5	0.6	0.1	223	0.0	0.0	0	0.0	0.0	0
10	0.5	0.1	223	0.0	0.0	0	0.0	0.0	0
15	0.4	0.0	223	0.0	0.0	0	0.0	0.0	0
Scenario 2									
5	2.8	0.2	465	0.0	0.0	0	0.0	0.0	0
10	1.9	0.1	465	0.0	0.0	0	−0.1	0.0	254
15	1.1	0.1	465	0.0	0.0	0	−0.1	0.0	333
Scenario 3									
5	6.0	0.2	499	0.0	0.0	0	−0.2	0.0	392
10	1.5	0.1	498	0.0	0.0	0	−0.6	0.0	499
15	0.0	0.0	55	0.2	0.0	230	−0.5	0.0	485
Scenario 4									
5	6.2	0.1	500	0.0	0.0	0	−0.6	0.0	422
10	0.0	0.0	32	0.7	0.1	339	−1.1	0.0	484
15	0.0	0.0	0	2.3	0.0	500	−0.1	0.0	15
Scenario 5									
5	0.9	0.1	283	0.2	0.1	89	−1.3	0.0	473
10	0.0	0.0	0	3.3	0.0	500	0.0	0.0	0
15	0.0	0.0	0	2.8	0.0	500	0.0	0.0	0

10%, or increase 20%. In the interest of space, we here provide results only on Setting 1 and note that the main conclusions under Setting 2 are similar, but the differences between the perturbed and original results are more pronounced. We also do not report our sensitivity results on the comparison of ReSTART policies with InvSTART here because the insights gained are very similar for those obtained on the comparison with START.

The results of the sensitivity analysis of comparison of ReSTART policies with START under Setting 1 are summarized in Table 3. This table follows the same

Table 3 Mean Reduction in the Critical Mortality Rate Obtained Using QD-ReSTART(0.5) and QS-ReSTART(0.5) Instead of START on Instances with Perturbed Reward Function Parameters

	QD-ReSTART vs. START				QS-ReSTART vs. START			
K	Mean (%)	HW (%)	No. of better instances	No. of worse instances	Mean (%)	HW (%)	No. of better instances	No. of worse instances
Scenario 1								
5	0.6	0.1	223	0	0.6	0.1	223	0
10	0.5	0.1	223	0	0.5	0.1	223	0
15	0.4	0.0	223	0	0.4	0.0	223	0
Scenario 2								
5	2.8	0.2	465	0	2.8	0.2	465	0
10	2.0	0.1	465	0	1.9	0.1	465	0
15	1.2	0.1	465	0	1.1	0.1	465	0
Scenario 3								
5	6.2	0.2	499	0	6.0	0.2	499	0
10	2.1	0.1	499	0	1.6	0.1	486	3
15	0.6	0.0	461	0	0.0	0.0	87	48
Scenario 4								
5	6.8	0.2	500	0	6.2	0.2	500	0
10	1.2	0.1	455	0	0.1	0.1	67	58
15	0.1	0.0	36	0	0.0	0.0	0	0
Scenario 5								
5	2.2	0.1	441	2	0.9	0.2	251	94
10	0.0	0.0	4	0	0.0	0.0	0	0
15	0.0	0.0	0	0	0.0	0.0	0	0

format as Table 2, except that we now separate the number of instances where ReSTART policies are statistically better from those in which they are statistically worse, both at a significance level of 0.05.

From this sensitivity analysis, we observed that while a few of the perturbed instances resulted in an increase in critical mortality, the vast majority of instances still saw improvement using QD-ReSTART. Even the first quartile had nonnegative improvement over both START and InvSTART in every scenario, and QD-ReSTART performed worse than START only in 2 (82) instances (out of the 7,500 simulated) at the 0.05 significance level under Setting 1 (2). QS-ReSTART has similar trends in average performance, but START performed better than QS-ReSTART in a larger number of instances (203 (281) of the 7,500 simulated under Setting 1 (2)) at the 0.05 significance level. Hence, QS-ReSTART appears to be less robust to perturbations or changes in the reward functions than QD-ReSTART. This difference can be attributed to the fact that under QD-ReSTART, a small change in the value of t_m due to imperfect information about reward functions will only lead to a correspondingly small change in t^*, which will not affect the policy significantly. On the other hand, under QS-ReSTART, a change in the value of t_m could lead to a complete reversal of the policy (from START to InvSTART or vice versa).

7. Adapting ReSTART to Changing Conditions on the Field

This section relaxes some of the fundamental assumptions that we made when building our mathematical and simulation models in the earlier sections. In particular, we make three structural changes to our original setup. First, because this paper focuses on incidents where no extensive search and rescue effort is needed, we originally assumed that all patients are accounted for immediately. However, even when no significant time is needed to locate and prepare patients for transportation, there could be still some delays in having some of the patients ready for transport. Hence, in this section, we will consider the possibility that not all patients are available at time zero and thus there is a delay in not only having these patients available but also knowing how many patients there are in total.

Second, we originally assumed that all patients are classified into the delayed and immediate categories correctly. In reality, triage is prone to errors. More specifically, there are two types of triage errors: *undertriage*, when a patient who should have been classified as immediate is classified as delayed, and *overtriage*, when a patient who should have been classified as delayed is classified as immediate. Taking these misclassification errors into the decision-making process is actually a subtle issue because it is tightly related to the estimation of the survival probabilities. Because classification errors are common in triage, survival probability functions can (and should) be estimated by taking this fact into account explicitly. Nevertheless, in practice, classification errors can be even more significant than the normally anticipated levels, and thus it is of interest to investigate their effect on the performance of prioritization policies. Thus, in this section, we will also consider the possibility that patients are misclassified as a result of triage.

Third, in connection with the second issue, our original model assumed that patients are not triaged again after time zero, which may be the case in practice because of lack of resources or poor organization. However, because many from the emergency response community emphasize the importance of *retriage*, and because our new set of relaxed conditions that allow misclassification make retriage a meaningful action, we now assume that patients will go through retriage some time after the start of the response effort.

The main difficulty that arises as a result of the explicit consideration of these new features is that there is a delay of information not only on the actual number of patients from each class but also on the total number of patients who will need to be transported. As one can observe from Figure 2, an incorrect estimate of the number of patients can change the specific prescription by ReSTART policies, which would possibly degrade their performance. In the remainder of this section, we discuss how one can use ReSTART policies in an adaptive way and test by means of a simulation study how their performance is affected by misclassification errors, delayed availability of patients, and retriage.

7.1. Adaptive QD-ReSTART and QS-ReSTART

Adaptive QD-ReSTART is essentially QD-ReSTART with policy parameters updated regularly or every time an event provides new information on the number of patients. This event could be the arrival of a new patient (which could change n_I, n_D, and $n_I + n_D$) or retriage (which could change n_I and n_D but not $n_I + n_D$). More specifically, adaptive QD-ReSTART uses the most up-to-date information on the number of patients in each class and determines which class should get a priority by following the QD-ReSTART description provided in §5 and using the τ value updated with respect to the current time t, which we call $\tau(t)$. In particular, we let $\tau(t) = t_m - t - n_D(t)\theta\phi/K$, where $n_i(t)$ is the number of patients categorized as class $i \in \{I, D\}$ at time t. (Note that $\tau(0)$ corresponds to τ in the description of QD-ReSTART provided in §5). Thus, $\tau(t)$ changes with the number of delayed patients, but not with the number of

immediate patients. Similarly, one could obtain an adaptive version of QS-ReSTART as follows: at time t, use START if $n_I(t) + n_D(t) \leq K(t_m - t)/(\phi\theta)$, and use InvSTART otherwise.

7.2. Performance of Adaptive ReSTART Policies

In this section, we report the results of a numerical study on the performance of the adaptive QD-ReSTART and QS-ReSTART policies. In this study, we used the same experimental setup as in §6.2 for the total number of patients, the number of ambulances, and the survival probability functions, which gave us the same 7,500 problem instances. However, this time, we also considered the possibility that there is some delay in having some of the patients ready for transportation and that the triage classification is imperfect. We assumed that the reward functions are the same (and have the same time origin at $t = 0$) for the patients who are ready to be transported and patients who are not ready. We also assumed that patients would go through triage once again at some point during the response effort.

We conducted two separate simulation studies. In Study 1, we assumed that all patients are available for transportation at time zero; however, each patient can be misclassified. Retrospective studies of mass-casualty incidents found that whereas the overtriage rate is typically high, the undertriage rate is usually very low. For example, Frykberg (2005) reported a range of 20%–80% for overtriage, whereas undertriage is almost nonexistent. To account for the relatively wide range of possible overtriage rates, we simulated each instance twice, once with a moderate overtriage probability for each delayed patient of 0.4, and another time with a high overtriage probability of 0.6. (We let p_O denote the overtriage probability.) We set the undertriage probability for each immediate patient to 0.05. For both cases, we assumed that patients who are still waiting for transportation 40 minutes after the incident go through retriage, which then reveals patients' true classifications. The choice of 40 minutes as the retriage time is largely arbitrary. In practice, emergency responders are urged to conduct retriage to identify cases of overtriage and undertriage (Mistovich and Karren 2007), but START and similar policies lack a standard dictating when retriage should occur (Lerner 2008).

We considered four different policies, namely, the adaptive versions of START, InvSTART, QD-ReSTART, and QS-ReSTART. Note that here START and InvSTART are adaptive in the sense that they use the updated classifications as a result of retriage. This is somewhat different from the notion of being adaptive for QD-ReSTART and QS-ReSTART, under which not only the classification of patients but also the structure of the policy may change due to retriage. Our results on the comparison of QD-ReSTART with START are summarized in Table 4 under the column labeled Study 1. Numbers given are the mean percentage improvement in the critical mortality rate when using the adaptive version of QD-ReSTART as

Table 4 Mean and 95% Half-Width for Reduction in the Critical Mortality Rate Obtained Using Adaptive ReSTART Policies Instead of START for Different Simulated Scenarios

	QD-ReSTART (Study 1) (%)				QD-ReSTART (Study 2) (%)				QS-ReSTART (Study 2) (%)			
	$p_0 = 0.4$		$p_0 = 0.6$		Moderate unav.		High unav.		Moderate unav.		High unav.	
K	Mean	HW	Mean	HW	Mean	HW	Mean	HW	Mean	HW	Mean	HW
Scenario 1												
5	0.5	0.1	0.5	0.1	0.5	0.1	0.3	0.0	0.5	0.1	0.3	0.0
10	0.4	0.0	0.3	0.0	0.4	0.0	0.2	0.0	0.4	0.0	0.2	0.0
15	0.3	0.0	0.2	0.0	0.3	0.0	0.1	0.0	0.3	0.0	0.1	0.0
Scenario 2												
5	2.5	0.1	2.3	0.1	2.4	0.1	1.9	0.1	2.4	0.1	1.9	0.1
10	1.6	0.1	1.4	0.1	1.6	0.1	1.1	0.1	1.6	0.1	1.1	0.1
15	0.9	0.1	0.8	0.0	0.9	0.0	0.6	0.0	0.8	0.0	0.4	0.0
Scenario 3												
5	5.6	0.2	5.3	0.2	5.6	0.2	4.9	0.1	5.5	0.2	4.8	0.1
10	1.7	0.1	1.6	0.1	1.7	0.1	1.6	0.0	0.7	0.1	0.2	0.0
15	0.5	0.0	0.4	0.0	0.4	0.0	0.6	0.0	0.0	0.0	0.0	0.0
Scenario 4												
5	6.3	0.1	6.0	0.1	6.2	0.1	5.7	0.1	6.1	0.1	5.2	0.1
10	1.1	0.0	1.1	0.0	1.1	0.0	1.4	0.0	0.0	0.0	0.0	0.0
15	0.1	0.0	0.1	0.0	0.1	0.0	0.5	0.0	0.0	0.0	0.0	0.0
Scenario 5												
5	2.4	0.1	2.4	0.1	2.4	0.1	2.5	0.1	0.7	0.1	0.2	0.1
10	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0
15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

opposed to START. Comparison with Table 1 reveals that with misclassification and retriage, the improvements by adaptive QD-ReSTART over START are slightly lower, but they are still significant. In the interest of space, we do not present here our results on the comparison of QD-ReSTART with InvSTART, which provided similar conclusions. We also omit our results on the performance of QS-ReSTART under Study 1 because the performance does not change much compared to the case without misclassification and retriage, which is reported in Table 2. This is an expected result because QS-ReSTART depends only on the total number of patients, and hence the policy structure does not change with misclassification errors.

In Study 2, we considered two cases where not all patients are available at time zero. In particular, under the *moderate unavailability* case, each patient has a probability 0.6 of being available at time zero, a probability 0.2 of being available at $t = 20$ minutes, and a probability 0.2 of being available at $t = 40$ minutes. For the *high unavailability* case, each patient has a probability 0.2 of being available at time zero, has a probability 0.4 of being available at $t = 30$ minutes, and a probability 0.4 of being available at $t = 60$ minutes. Initial triage is assumed to have been done immediately after each arrival, and a retriage is carried out once all patients are available. We also assume a moderate overtriage probability of 0.4 and a low undertriage probability of 0.05. The decision maker does not have advance knowledge of how many patients there are in total and the times at which new patients are going to become available. Therefore, idling to wait for a higher-priority patient is not considered as long as there are patients in need of transport.

Results on the comparison of QD-ReSTART and QS-ReSTART with START under Study 2 are provided in Table 4. These results show that although the performance improvement with adaptive ReSTART policies is smaller than in the case where all patients are available at time zero, it is still significant, especially in Scenarios 2–4 with a low level of resources. As expected, the performance improvement decreases when more patients are unavailable at time zero and they are unavailable for a longer period of time.

8. Conclusions

With this work, we have demonstrated that it is possible to design a prioritization policy that takes into account the three main components of the mass-casualty triage problem (i.e., the size of the event, availability of resources, and dependence of survival probabilities on time) in a very simple way and performs better, substantially at times, than the common practice (START) that largely ignores these components. In particular, using a fluid formulation, we identified characteristics of "good" resource-based prioritization policies, which led to a simple policy that we call ReSTART and its variations. Using realistic simulations with data from emergency medicine literature, we have observed that these policies have the potential to improve the critical mortality rate over START.

It would be naive to claim that ReSTART policies are readily available for implementation in the exact same way we described them here. Any such policy that would radically change the adopted practice needs to be scrutinized carefully by the medical community before being formally proposed as an alternative, and such scrutiny may necessitate some adjustments. Nevertheless, we believe that the structural properties of ReSTART provide insights that can be useful in efforts to develop resource-based prioritization policies in practice. For example, we have observed that it is best to follow START (i.e., to give priority to immediate patients at all times) when there are *few patients* compared with the number of available resources and best to follow InvSTART (i.e., to give priority to delayed patients at all times) when there are *many delayed patients*. Otherwise, it is best to use a policy that prioritizes immediate patients initially but switches to delayed patients at *some point in time*. ReSTART gives a precise description of this structure by quantifying what *few patients*, *many delayed patients*, and *some point in time* really mean. Even if practitioners do not follow this description exactly, they can still build another policy having a structure that is similar to that of ReSTART, perhaps by coming up with new definitions for what it means to have few patients or many delayed patients. In short, our analytical characterization can provide a broad outline for the type of policy that is expected to work well in practice.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/msom.1120.0426.

Acknowledgments

The authors thank James E. Winslow for providing valuable advice on medical aspects of this paper. The authors also thank the editor-in-chief, the associate editor, and three anonymous referees who provided comments that significantly improved this paper. This work was partially supported by the National Science Foundation [Grants CMMI–0927607, CMMI–1234212].

References

Argon NT, Winslow JE, Ziya S (2011) Triage in the aftermath of mass-casualty incidents. Cochran JJ, Cox LA, Keskinocak P, Kharoufeh JP, Smith JC, eds. *Wiley Encyclopedia of Operations Research and Management Science* (Wiley, Hoboken, NJ), 5611–5620.

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.

Baker R, Strosberg M (1992) Triage and equality: An historical reassessment of utilitarian analyses of triage. *Kennedy Institute of Ethics J.* 2(2):103–123.

Cone DC, MacMillan DS (2005) Mass-casualty triage systems: A hint of science. *Acad. Emergency Medicine: Official J. Soc. Acad. Emergency Medicine* 12(8):739–741.

Cox DR, Oakes D (1984) *Analysis of Survival Data* (CRC Press, London).

Frykberg ER (2005) Triage: Principles and practice. *Scandinavian J. Surgery* 94(4):272–278.

Garner A, Lee A, Harrison K, Schultz CH (2001) Comparative analysis of multiple-casualty incident triage algorithms. *Ann. Emergency Medicine* 38(5):541–548.

Ingolfsson A, Budge S, Erkut E (2008) Optimal ambulance location with random delays and travel times. *Health Care Management Sci.* 11(3):262–274.

Jenkins JL, McCarthy ML, Sauer LM, Green GB, Stuart S, Thomas TL, Hsu EB (2008) Mass-casualty triage: Time for an evidence-based approach. *Prehospital Disaster Medicine* 23(1):3–8.

Kennedy K, Aghababian R, Gans L, Lewis CP (1996) Triage: Techniques and applications in decision making. *Ann. Emergency Medicine* 28(2):136–144.

Lerner EB (2008) Mass casualty triage: An evaluation of the data and development of a proposed national guideline. *Disaster Medicine Public Health Preparedness* (1):S25–34.

Mistovich JJ, Karren KJ (2007) *Prehospital Emergency Care*, 8th ed. (Prentice Hall, Upper Saddle River, NJ).

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

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

Winslow GR (1982) *Triage and Justice* (University of California Press, Berkeley).

Winslow JE (2010) Personal communication with Alex Mills via email, May 10.