IIE Transactions (2014) **46**, 728–741

ISSN: 0740-817X print / 1545-8830 online
DOI: 10.1080/0740817X.2013.855846

Minimizing mortality in a mass casualty event: fluid networks in support of modeling and staffing

IZACK COHEN*, AVISHAI MANDELBAUM and NOA ZYCHLINSKI

Technion, Industrial Engineering and Management, Haifa, Israel
E-mail: izik68@tx.technion.ac.il

Received November 2012 and accepted August 2013

The demand for medical treatment of casualties in mass casualty events (MCEs) exceeds resource supply. A key requirement in the management of such tragic but frequent events is thus the efficient allocation of scarce resources. This article develops a mathematical fluid model that captures the operational performance of a hospital during an MCE. The problem is how to allocate the surgeons—the scarcest of resources—between two treatment stations in order to minimize mortality. A focus is placed on casualties in need of immediate care. To this end, optimization problems are developed that are solved by combining theory with numerical analysis. This approach yields structural results that create optimal or near-optimal resource allocation policies. The results give rise to two types of policies, one that prioritizes a single treatment station throughout the MCE and a second policy in which the allocation priority changes. The approach can be implemented when preparing for MCEs and also during their real-time management when future decisions are based on current available information. The results of experiments, based on the outline of real MCEs, demonstrate that the proposed approach provides decision support tools, which are both useful and implementable.

Keywords: Mass casualty events, fluid models, resource allocation, optimal policy

1. Introduction

Mass Casualty Events (MCEs) occur quickly and suddenly. They produce a relatively large number of casualties who need immediate care and thus overwhelm hospital resources. They frequently occur due to terror attacks, accidents, or natural disasters. For example, on the morning of July 7, 2005, terrorists launched a series of attacks across London that left 56 people dead and 775 injured (Aylwin *et al*., 2006); a Buenos Aires train crashed in 2012, resulting in more than 700 injuries (BBC News, 2012); and our partner hospital has, unfortunately, gathered ample experience in catering to MCEs after terror events—an experience that will guide us later on in our examples.

The environment of an Emergency Department (ED) in a hospital during an MCE is stressful. People run around frantically, casualties' cries emanate from the treatment rooms, and worried relatives hope for encouraging news. During this time it is imperative to deliberately manage the event and make, as far as possible, the right clinical and operational decisions. Figure 1 illustrates the flow of casualties through a hospital after an MCE. (Our showpiece for this article is a large Israeli hospital that has become experienced in handling emergencies; our modeling framework nevertheless is general and can be easily modified to accommodate other hospitals.)

On arrival, casualties are triaged and prioritized for treatment according to their medical situation (Mehta, 2006). There are several triage systems that distinguish between several classes of casualties (e.g., Lerner *et al*. (2008)). Our hospital uses a simple in-hospital triage system that classifies arriving casualties according to one of two categories: Immediate or Not Immediate. We focus on the former category: it concerns casualties who are in danger of dying unless provided with prompt medical treatment; this entails stabilizing life-saving treatment and for some also an immediate operation, which underscores the significance of appropriately managing medical resources.

After an MCE, there is a mounting demand for medical treatment, typically far in excess of the existing capacity to administer it. Consequently, the medical staff, and especially the surgeons who are most frequently the bottleneck resource (Hirshberg *et al*., 1999; Einav *et al*., 2006), cannot provide prompt treatment to all casualties. Casualties classified as Immediates are prioritized for prompt treatment. However, it may turn out to be impossible to attend promptly to all the Immediates; therefore, the main objective of MCE management is to minimize their mortality.

*Corresponding author
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uiie.

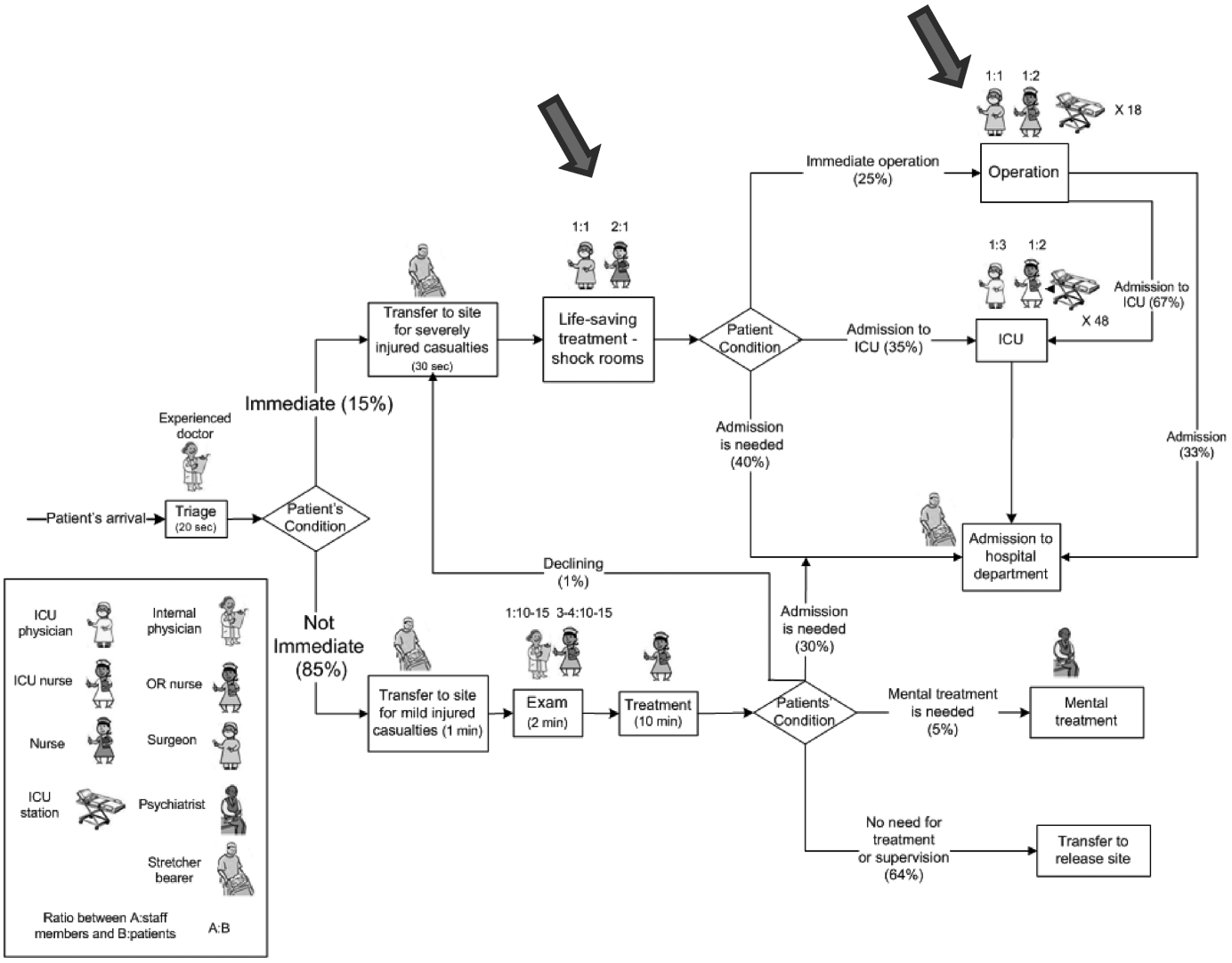

Fig. 1. An illustration of the flow of casualties through an ED during an MCE. The tilted arrows mark the two stations upon which we focused.

In this article we develop allocation policies of surgeons to two treatment stations (marked by the tilted arrows in Fig. 1), which seek to minimize the overall mortality of immediate casualties during an MCE.

Our model can be used during preparations for MCEs, as well as for supporting their real-time management. For preparations, the resource allocation policies are based on estimates of reference scenario parameters, such as treatment and mortality rates at a station. For real-time MCE management the proposed model exploits data that are continuously being updated as the event unfolds, such as changes in the initial forecast given for the arrival of casualties. It then solves a rolling horizon resource allocation problem.

Our work can be considered as the analysis of a two-stage tandem queueing system with flexible servers, customer abandonments, and time-varying arrivals; the literature on tandem queueing systems, however, is mostly focused on steady-state without abandonments (Ahn *et al*., 2002; Andradottir and Ayhan, 2005). Our approach also differs from most operations management research on MCEs, which commonly uses simulations for modeling and evaluation of alternative management policies (Hirshberg *et al*., 1999; Sinreich and Marmor, 2004; Paul *et al*., 2006). Although the benefit of complex systems simulation cannot be underestimated, it provides limited structural insights, and conceivably cannot (yet) support real-time management of MCEs. We follow the emerging stream of research whereby fluid models support healthcare operations management (Yom-Tov, 2007; Argon *et al*., 2008). We thus propose a fluid modeling framework that is suitable for capturing the transient finite-horizon evolution (as opposed to steady-state) of MCEs. This framework calls for a focus on time-inhomogeneous predictable variability, which fluid models ideally capture. Thereafter, we use our framework to identify structural properties of stylized fluid models. These

structural properties yield management policies that minimize the number of fatalities. The suggested policies are both insightful and implementable.

The article is organized as follows. We review the relevant literature in the next section. Section 3 describes MCE environments, modeling assumptions, and the model formulation. Section 4 contains an analysis of optimal resource allocation policies and provides managerial insights for applying them. In Section 5, we apply and test our results against the outline of two MCEs. Section 6 extends the model to cases where resource allocation decisions are made periodically and to real-time management of MCEs. The final section offers worthy directions for further research.

2. Literature review

Our approach follows two streams of research: MCEs and fluid models. In this section we briefly review the relevant literature from both streams and the one example we found that combines the two. The MCE-related literature is diverse: it analyzes clinical (Hirshberg *et al*., 2001; Hirshberg *et al*., 2005; Aylwin *et al*., 2006), social science (Hughes, 1991; Altay and Green, 2007; Merin *et al*., 2010), and operational aspects. We focus on operational aspects for which the relevant research is limited (Altay *et al*., 2007) and the problems are challenging, even when compared with the clinical aspects (Waeckerle, 1991).

When an MCE occurs the Immediates are treated first (Lerner, 2008). The main objective is then to reduce the mortality of its treated casualties by providing them with a "level of care that approximates the care given to similar casualties under normal conditions" (Hirshberg *et al*., 2001; Hirshberg *et al*., 2005, p. 647).

With that in mind, we seek to develop resource management policies that minimize the mortality of Immediates during an MCE. The need for such policies becomes clear from their in-practice application: one example is the Israeli field hospital in Haiti that was established after the January 2010 earthquake; it treated approximately 100 casualties per day and had a capacity of 60 beds that was later increased to 72 beds. Kreiss *et al*. (2010) and Merin *et al*. (2010) reported that dynamic resource allocation and staffing enhanced the efficiency of that hospital.

Several researchers have dealt with resource allocation during MCEs. Argon *et al*. (2008) developed state-dependent heuristic prioritization policies for casualties being treated by a single-server clearing system. Casualties who are not treated within their "lifetime" die and the objective is to maximize the expected number of survivors. Jacobson *et al*. (2012) extended the latter research to consider different mortality probabilities for different types of casualties and multiple resources. Both these models assume that all casualties are available at the outset of the MCE, triaged to different priority categories, and that there is a single station. Mills *et al*. (2013) examined a possible scenario for these models—the evacuation by ambulances of casualties from an MCE arena to a hospital. They developed prioritization policies for different casualty classes that were triaged *in situ*. Our model's focus is on the ED arena where casualties arrive continuously, and the surgeons are required to be in two different places "at once" in order to take care of a single type of casualties (i.e., Immediates). Our choice of surgeons as the scarce resource is supported by the experience of our hospital partners, as well as by Einav *et al*. (2006). The latter collected their data at trauma centers in Israel from 32 MCEs caused by suicide bombings. Their analysis indicates that the surgeons represent a scarce resource that is needed in the ED and the operating rooms simultaneously.

Our model also contributes to the almost non-existent literature about designing the surge capacity of a hospital. Hick *et al*. (2004, p. 254) defined surge capacity as the "ability to manage a sudden, unexpected increase in patient volume (i.e., number of patients) that would otherwise severely challenge or exceed the current capacity of the health care system." Examples of fundamental questions that must be addressed when designing surge capacity are: how many casualties are expected at the different treatment stations concurrently, and what is the estimated time from the start of the event until the peak demand at these stations. The definition of what constitutes surge capacity varies; it typically follows rules of thumb such as, when determining surge capacity by the percentage of the hospital's bed capacity (e.g., the Israeli Ministry of Health sets a hospital's surge capacity at 20% of its beds). At other times, surge capacity is set according to a fixed number of casualties based on past events (Kosashvili *et al*., 2009), simulations of performance as a function of the casualties arrival rate (Hirshberg *et al*., 2005), or the time between the start of an MCE until the trauma teams reach their full capacity (Hirshberg *et al*., 2010). As a by-product of our approach, which finds the best resource allocation policies to minimize mortality, we forecast the time and magnitude of the peak demand at the treatment stations. Our model can thus be used to support decisions for designing surge capacity by performing a sensitivity analysis on the level of resources, consequently estimating the allowed time from the start of an event to when an increase in capacity is needed (e.g., recruit surgeons from the hospital or, alternatively, direct casualties to other hospitals).

Simulation is widely accepted as an effective method for assisting management in healthcare decision making. The simulation model of Sinreich and Marmor (2004) is an excellent example of this approach. It was developed for short-term operational planning in EDs. Based on data from 12 urban terrorist bombing events, Hirshberg *et al*. (1999) developed a simulation model of an ED during such events. They concluded that the surge capacity of a hospital depends primarily on the number of available surgeons; they then defined an optimal staff profile for surgeons and

trauma nurses that arise as the scarce resources. Paul *et al*. (2006) used simulations for predicting casualties' waiting times, and for estimating hospitals' capacities within a disaster region.

We use fluid models that account for the transient nature of MCEs. We adopt this approach because of its analytical tractability, which leads to optimal policies for simple, yet realistic, MCE scenarios. In such models, the entities that move through the system (e.g., casualties) are assumed to be fluid and so the flow can be described through differential equations. The literature indicates that fluid models (or approximations) are accurate for heavily loaded service systems. For example, Mandelbaum, Massey, Reiman, and Rider (1999) and Mandelbaum, Massey, Reiman, and Stolyer (1999) developed fluid approximations for a multiserver single queue with abandonment and retrials. The model was proven accurate both in its steady state and in its transient state; the latter was caused by a sudden peak in the casualties' arrival rate, as is typical during an MCE. Mandelbaum, Massey, Reiman, and Stolyer (1999) showed that waiting time approximations are asymptotically exact as the size of the system increases. Our model exploits performance measures, such as the time-varying number of people in the system and the number at each station, which allow one to develop resource allocation policies. Fluid models of service systems have been extended to include state-dependent arrival rates and general arrival and service rates (Whitt, 2005, 2006).

Fluid models have been successfully implemented in different types of service systems. These cover the early applications for post offices, claims processing in a Social Security office (Oliver and Samuel, 1962; Vandergraft, 1983), and more recently a financial service call center (Green *et al*., 2005).

The research setting of Yom-Tov (2007) is perhaps the closest to ours. She developed fluid and diffusion limits for the Erlang-R model, which accommodates returns of customers to service. These limits lead to fluid approximations that are not only useful in analyzing time-varying systems, but they also help understand their transient behavior. Her model was used to analyze MCEs in which the arrival rate changes rapidly during a short period of time. A numerical example in which the arrival rate is multiplied fivefold over 2 hours was simulated and compared with its fluid and diffusion approximations. The comparison demonstrated a high degree of accuracy.

Our model differs from the existing literature in two fundamental ways: first, it deals with a situation in which the casualties arrive at a hospital according to a general arrival rate and, second, we explicitly consider two stations, in tandem, where medical treatment is delivered by the same scarce resource.

Our primary focus is to minimize the number of mortalities, and so we seek resource allocation solutions and policies for planning and real-time management of MCEs. Note that prior research, which dealt with our problem specification, presents results of simulations and numerical analysis that can be computationally intensive and provides limited general insight.

3. The model

This section starts with a discussion of the environment, the assumptions that we make, and the dynamics during an MCE. In subsection 3.2 we introduce notations and formulate the problem.

3.1. *The model's environment, assumptions, and dynamics*

We model part of an ED in a hospital during an MCE. As explained, we presume that surgeons are the bottleneck resource during MCEs.

Our fluid model approximates ED dynamics during an MCE. Casualties arrive at the ED continuously (e.g., a given reference scenario). If there are enough resources, then casualties are admitted for treatment, which is either life-saving (Station 1) or an operation (Station 2). These are performed at a known service rate. Casualties may die either during treatment or while waiting for it. The mortality rates can be different at different stations. Mortality rates can be interpreted as either fatality rates of casualties or as operational constraints. In the latter case, they are the reciprocals of the average maximum allowable time for a casualty to complete treatment, in order to avoid fatality (Paul *et al*., 2006). We assume that mortality rate is constant for each station, which enables explicit analytical solutions and structural insights and, equally important, it is reasonable since it does in fact capture the underlying stochastic death times—some long, others short. Our models can nevertheless accommodate differing rates, or merely constraints on waiting times, albeit at the cost of insight and tractability. One could also argue in favor of stochastic dependence of death times across stations, but this would lead to far more complicated models (Pang and Whitt, 2012), which we leave for future research.

As common in practice, we assume that one surgeon treats a single casualty at either one of the stations (Hirshberg *et al*., 1999; Aylwin *et al*., 2006). It is worth noting that one surgeon may treat several mortal-risk casualties at the same time in response to a specific real-time crisis. However, this is undesirable and does not change the medical policies; therefore, we do not take such an option into account. We believe that the preparation for an MCE should be based on a model that takes the standard medical practices into consideration. Moreover, our belief is that real-time emergency decision making may relax some of the assumptions that we made during the preparation phase, as a prompt solution to a local crisis. There is also a technical reason for assuming the constant casualty–surgeon ratio—it facilitates the model's formulation; changing it

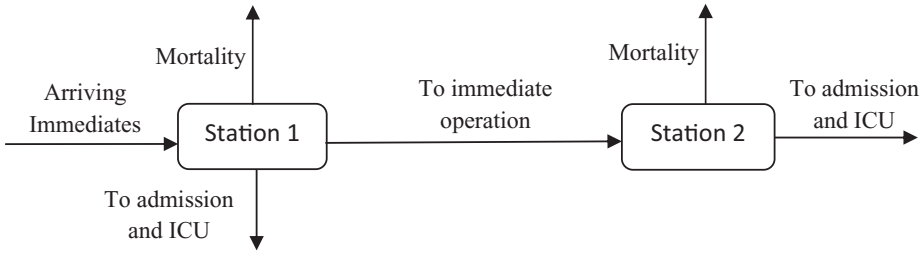

Fig. 2. Casualties flow in a two-station network.

will significantly complicate the model and the structural results that we achieve.

The routing probabilities and the duration of treatments in Fig. 1 were estimated by expert trauma doctors. Nevertheless, our modeling framework is general and we analyze different scenarios. The tilted arrows in Fig. 1 mark the two stations on which we focus, while Fig. 2 introduces the corresponding basic stylized model, which captures the conflict of surgeon allocation.

The sequence of events for Fig. 2 is as follows:

1. Immediate casualties arrive at Station 1 after being triaged. We assume that all Immediates share similar (severe) clinical assessment. Their flow through the network would therefore be according to first-come first-served priority.
2. An Immediate who enters Station 1 at time t receives life-saving treatment if at least one of the $N_1(t)$ surgeons is available; if not, she or he waits for treatment.
3. With probability p_{12}, the casualty who finished treatment at Station 1 is directed to Station 2, where $N_2(t)$ surgeons are allocated. An available surgeon starts treatment immediately; alternatively, the casualty must wait for a surgeon who is in the middle of treatment.
4. Treatment rates are μ_1 and μ_2 for Stations 1 and 2, respectively.
5. Casualties may die either while waiting for or receiving treatment. The mortality rates are θ_1 and θ_2 for Stations 1 and 2, respectively. A reasonable assumption is that $\theta_1, \theta_2 \ll 1$ (time units throughout the article are in minutes).
6. The "effective" treatment time for a casualty treated at a station includes the duration of the treatment and any time delay caused by unavailable surgeons.

Treatment may take place at a station only if the necessary resources (e.g., surgeons, operating rooms, and medical equipment) are available. We assume that the only constraining resources are the N surgeons who are available at the hospital, or formally $N_1(t) + N_2(t) \leq N$, at all times $t \geq 0$ during the MCE.

Our key technique is to prepare for an MCE by assuming a reference scenario and finding the best decisions for the surgeon allocations. These decisions impact the waiting times of casualties, their flow through the network, and their likelihood of survival. Our fluid model approach captures the dynamic nature of an MCE and suggests dynamic policies; these may be ignored if we either assume time-homogenous parameters (e.g., constant arrival rate) or use a steady-state model (e.g., steady-state simulation or queueing theory approximations).

The fluid model serves as an approximation of the underlying stochastic environment in which arrivals, mortality, treatment times, and the other parameters are random variables. Therefore, in addition to the support from the literature that fluid models should provide good approximations of their corresponding stochastic environment, we conducted experiments to validate the accuracy of our fluid model when used to capture our specification of the problem as presented in Fig. 2. These experiments, each using 500 simulation replications, compared the fluid model results against a discrete-event stochastic simulation in which casualties arrive according to a non-homogenous Poisson process that was used to represent a general, time-dependent arrival rate; treatment durations were randomly generated from exponential distributions. In all cases, the fluid model forecasted, rather accurately, the stochastic behavior of the corresponding simulation in that its results were nearly always within the bounds of the 95% simulation confidence interval and always within the (wider) bounds of a 99% confidence interval (Fig. 3 is a representative example).

3.2. *Model formulation*

Table 1 includes the notations used throughout the article. Whenever possible we suppress subscripts to improve readability.

We assume that by the time the first casualty arrives at the hospital all prior casualties will have been cleared. We then choose T to be large enough to ensure that the last casualty has completed treatment by time T; $[0, T]$ is hence the time interval over which our model is formulated.

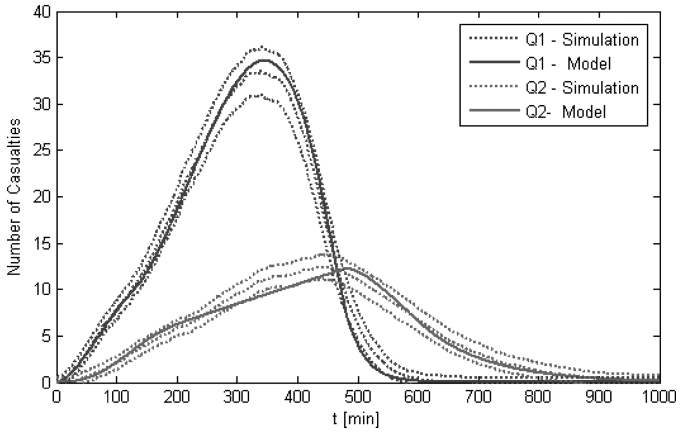

Fig. 3. The number of casualties in Stations 1 and 2 (Q1 and Q2) as a function of time—simulation and fluid model results. The upper and lower dotted lines are, for each station, the upper and lower limits of a 95% confidence interval respectively, and the solid lines correspond to the fluid model.

First we introduce a continuous optimization problem with its objective to minimize total mortality:

$$\min_{N_1(\cdot),N_2(\cdot)} \int_0^T [\theta_1 Q_1(t) + \theta_2 Q_2(t)]\mathrm{d}t,$$

such that for all $t \in [0, T]$:

$$\dot{Q}_1(t) = \lambda(t) - \mu_1(Q_1(t) \wedge N_1(t)) - \theta_1 Q_1(t),$$
$$\dot{Q}_2(t) = p_{12}\mu_1(Q_1(t) \wedge N_1(t)) - \mu_2(Q_2(t) \wedge N_2(t)) - \theta_2 Q_2(t),$$
$$N_1(t) + N_2(t) \le N,$$
$$N_1(t),\ N_2(t),\ Q_1(t),\ Q_2(t) \ge 0,$$
$$Q_1(0) = 0,\ Q_2(0) = 0.$$

Next we approximate the above problem by discretizing time. The discrete-time formulation has two advantages. The problem can be transformed, as we show in the following, into a standard Linear Program (LP), which is easy to solve even for a large number of variables and constraints.

Table 1. Notation

Variable	*Definition*
i	Identifies a treatment station
$\lambda(t)$	Arrival rate at Station 1 at time t
$N_i(t)$	Number of surgeons at Station i at time t
N	Number of available surgeons
S	Minimal time window for changing resource allocations
μ_i	Treatment rate at Station i per surgeon
θ_i	Mortality rate at Station i
$Q_i(t)$	Number of casualties at Station i (in queues and in treatment) at time t

The second advantage is that the discrete-time formulation can naturally accommodate constraints, which are linked to discrete time points (e.g., change resource allocations only every 30 minutes). Naturally, as the discrete time step gets smaller the approximation gets better:

(P1):

$$\min_{N_1(\cdot),N_2(\cdot)} \sum_{t=0}^{T} [\theta_1 Q_1(t) + \theta_2 Q_2(t)],$$
such that for $t = 0, 1, \ldots, T-1$:
$$Q_1(t+1) = Q_1(t) + \lambda(t) - \mu_1(Q_1(t) \wedge N_1(t)) - \theta_1 Q_1(t),$$
$$Q_2(t+1) = Q_2(t) + p_{12}\mu_1(Q_1(t) \wedge N_1(t)) - \mu_2(Q_2(t) \wedge N_2(t)) - \theta_2 Q_2(t),$$
$$N_1(t) + N_2(t) \le N,$$
$$N_1(t),\ N_2(t),\ Q_1(t),\ Q_2(t) \ge 0,$$
$$Q_1(0) = 0,\ Q_2(0) = 0.$$

For concreteness, we assume that the length of a time interval is 1 minute.

Problem (P1) is not linear due to the minimum between the number of casualties at a station and the number of surgeons who treat them. We propose an equivalent constraint formulation that is linear. The new formulation of Problem (P2) forces the number of surgeons at a station to be smaller or equal to the number of casualties at that station and avoids redundant allocations.

Proposition 1. *The optimal solutions for Problem (P1) are the same as those for the following Problem (P2):*

(P2):

$$\min_{N_1(\cdot),N_2(\cdot)} \sum_{t=0}^{T} [\theta_1 Q_1(t) + \theta_2 Q_2(t)],$$
such that for $t = 0, 1, \ldots, T-1$:
$$Q_1(t+1) = Q_1(t) + \lambda(t) - \mu_1 N_1(t) - \theta_1 Q_1(t),$$
$$Q_2(t+1) = Q_2(t) + p_{12}\mu_1 N_1(t) - \mu_2 N_2(t) - \theta_2 Q_2(t),$$
$$N_1(t) \le Q_1(t),$$
$$N_2(t) \le Q_2(t),$$
$$N_1(t) + N_2(t) \le N,$$
$$N_1(t),\ N_2(t),\ Q_1(t),\ Q_2(t) \ge 0,$$
$$\mathrm{Q}_1(0) = 0,\ \mathrm{Q}_2(0) = 0.$$

(Proofs for the propositions are provided in the Online Supplement.)

Proposition 1 not only helps mathematically by turning the formulation linear but it also frees up any redundant resources by forcing the allocations to be tight, which increases the operational flexibility of the decision maker.

Some additional mathematical manipulations yield LP (P3), which is more amenable for analysis. Note that in all of our formulations, there is an underlying assumption

that resource allocations of surgeons can be changed at any time, thus preempting treatments. This assumption is reasonable for MCEs that operate under heavy traffic regimes for which it has been proved (e.g., Atar *et al*., (2004)) that a non-preemptive policy is asymptotically equivalent to its preemptive counterpart. In reality it is common that allocation updates follow periodic assessments (e.g., every 30 minutes). In Section 6 we deal with such constraints and analyze their impact.

Proposition 2. *The formulation of Problem (P2) is equivalent to the following formulation (P3):*

(P3):

$$\min_{N_1(\cdot),N_2(\cdot)} \sum_{t=1}^{T-1} \{N_1(t)\mu_1[(1-\theta_1)^{T-t}-1 - p_{12}[(1-\theta_2)^{T-t}-1]] + N_2(t)\mu_2[(1-\theta_2)^{T-t}-1]\},$$

subject to

$$N_1(1) = 0,$$

$$\mu_1 N_1(1) + N_1(2) \le \lambda(1),$$

$$(1-\theta_1)\mu_1 N_1(1) + \mu_1 N_1(2) + N_1(3) \le (1-\theta_1)\lambda(1) + \lambda(2),$$

$$\vdots$$

$$(1-\theta_1)^{T-3}\mu_1 N_1(1) + (1-\theta_1)^{T-4}\mu_1 N_1(2) + \cdots + N_1(T-1) \le (1-\theta_1)^{T-3}\lambda(1) + (1-\theta_1)^{T-4}\lambda(2) + \cdots + \lambda(T-1),$$

$$N_2(1) = 0,$$

$$\mu_2 N_2(1) - \mathrm{p}_{12}\mu_1 N_1(1) + N_2(2) \le 0,$$

$$(1-\theta_2)\mu_2 N_2(1) - (1-\theta_2)\mathrm{p}_{12}\mu_1 N_1(1) + \mu_2 N_2(2) - \mathrm{p}_{12}\mu_1 N_1(2) + N_2(3) \le 0,$$

$$\vdots$$

$$(1-\theta_2)^{T-3}\mu_2 N_2(1) + (1-\theta_2)^{T-3}\mathrm{p}_{12}\mu_1 N_1(1) + (1-\theta_2)^{T-4}\mu_2 N_2(2) - (1-\theta_2)^{T-4}\mathrm{p}_{12}\mu_1 N_1(2) + \cdots \cdots + \mu_2 N_2(T-2) - \mathrm{p}_{12}\mu_1 N_1(T-2) + N_2(T-1) \le 0,$$

$$N_1(t) + N_2(t) \le N, \quad t = 0, 1, \ldots, T-1,$$

$$N_1(t),\ N_2(t) \ge 0, \quad t = 0, 1, \ldots, T-1.$$

Problem (P3) is solved by standard LP techniques (we used Matlab and Mosek toolbox; www.mosek.com) to find optimal dynamic surgeon allocations. In the next section we analyze the problem to identify structural properties of optimal policies.

4. Problem analysis

When surgeons are overloaded, the decision maker must prioritize them to either Station 1 or Station 2. The optimal allocation policy can be dynamic, in which priorities change (e.g., first prioritize Station 1 and at some later time prioritize Station 2), or static, in which priorities are kept fixed throughout the event.

In this section we characterize optimal priority settings for various scenarios, which are then followed by an analysis of each scenario. We start by introducing a greedy formulation for the optimization problem. This greedy formulation identifies, at each discrete time point $t = 0, 1, \ldots, T-1$ (t is the starting time of interval $t+1$), the surgeon allocations $N_1(t)$, $N_2(t)$ that minimizes mortality over interval $t+1$ only. To this end, we formulate a sequence of continuous Knapsack problems (Kellerer *et al*., 2004, pp. 17–20), indexed by $t = 0, 1, \ldots, T-1$. Problem t corresponds to time interval t, and it uses the quantities $\lambda(t)$, $Q_1(t)$, $Q_2(t)$ that are known at the end of interval $t-1$. Formally, for $t = 0, 1, \ldots, T-1$, the problem is

(P4):

$$\max_{N_1(t),N_2(t)} N_1(t)\mu_1[\theta_1 - p_{12}\theta_2] + N_2(t)\mu_2\theta_2,$$

subject to

$$N_1(t) \le Q_1(t),$$

$$N_2(t) \le Q_2(t),$$

$$N_1(t) + N_2(t) \le N,$$

$$N_1(t),\ N_2(t),\ Q_1(t),\ Q_2(t) \ge 0.$$

Proposition 3 characterizes the optimal allocation policy for the greedy Problem (P4). (Its proof is provided in the Online Supplement.)

Proposition 3. *An optimal policy for the greedy problem (P4) is to allocate to Station i^* all of the surgeons it requires from the N available, where $i^* = 1$ if $\mu_1(\theta_1 - p_{12}\theta_2) \ge \mu_2\theta_2$ and $i^* = 2$ otherwise; if there are still available surgeons left then allocate them to the other station.*

In other words, the prioritized (higher priority) Station i is allocated all its needed resources, to the extent possible: $\min(Q_i(t), N)$ surgeons; any remaining surgeons are assigned to the other station.

Note that for equal mortality rates, Proposition 3 determines station priorities according to the relative values of $\mu_1(1-p_{12})$ and μ_2. The greedy policy plays an important role in solving our original problem, as formulated in Problems (P1) to (P3). This role emerges by identifying nine cases, according to all possible combinations between $\mu_1(1-p_{12})$ and μ_2, θ_1 and θ_2.

In the first three cases, mortality rates are equal. Case 1: $\theta_1 = \theta_2$ and $\mu_1(1-p_{12}) = \mu_2$; Case 2: $\theta_1 = \theta_2$ and $\mu_1(1-p_{12}) > \mu_2$; Case 3: $\theta_1 = \theta_2$ and $\mu_1(1-p_{12}) < \mu_2$. For the next three cases the mortality rate is higher at Station 1. Case 4: $\theta_1 > \theta_2$ and $\mu_1(1-p_{12}) = \mu_2$; Case 5: $\theta_1 > \theta_2$ and $\mu_1(1-p_{12}) > \mu_2$; Case 6: $\theta_1 > \theta_2$ and $\mu_1(1-p_{12}) < \mu_2$. For the last three cases Station 2 has a higher mortality rate, Case 7: $\theta_1 < \theta_2$ and $\mu_1(1-p_{12}) = \mu_2$; Case 8: $\theta_1 < \theta_2$ and $\mu_1(1-p_{12}) > \mu_2$; Case 9: $\theta_1 < \theta_2$

Table 2. Summary of suggested priority settings of surgeon allocations

Conditions	$\theta_1 = \theta_2$	$\theta_1 > \theta_2$	$\theta_1 < \theta_2$
$\mu_1(1-p_{12}) = \mu_2$	Station 1 or 2—equal performance (Case 1)	Station 1 (Case 4)	Station 2 (Case 7)
$\mu_1(1-p_{12}) > \mu_2$	Station 1 (Case 2)	Station 1 (Case 5)	Prioritize Station 1 and switch priorities at some t (Case 8)
$\mu_1(1-p_{12}) < \mu_2$	Station 2 (Case 3)	Prioritize Station 2 and switch priorities at some t (Case 6)	Station 2 (Case 9)

and $\mu_1(1-p_{12}) < \mu_2$ (Table 2 lists these cases and their suggested priority settings, which we develop in the following discussions).

It turns out that the sequence of greedy solutions, via Problem (P4), in fact solves Problem (P3) when the mortality rates at both stations are equal (i.e., Cases 1 to3). Formally:

Proposition 4. *Assume that $\theta = \theta_1 = \theta_2$. Then an optimal solution of Problem (P3) is given by any sequence of greedy solutions for Problem (P4).*

The proof of this last proposition is rather tedious; hence, it is placed in the Online Supplement. The proof yields a static priority rule for an optimal surgeons' allocation.

We now extend Proposition 4 to some cases of unequal mortality rates, which we formalize as Proposition 5.

Proposition 5. *If Station i gets priority when $\theta_i = \theta_j$, then it will get priority when $\theta_i > \theta_j$.*

The details of the induction-based proof appear in the Online Supplement. The proposition explicitly identifies optimal allocation policies for Cases 4, 5, 7, and 9.

We are left with two more cases to consider—Cases 6 and 8. For these two cases, we have no closed-form analytical solution for the optimal policy. However, extensive numerical experiments suggest that, in these cases, the optimal policy switches station priority at some point in time, and in all experiments there was only a single such switch time. For Case 6, priority is first given to Station 2 and at some time switches to Station 1. Case 8 is the opposite: Station 1 is prioritized first, and at some later time priority is given to Station 2. Note that, in both cases, the priority switch point can be found by merely solving an LP optimization problem. Nevertheless, it is of interest to provide insights about the differences between greedy non-switching policies and the optimal policies for Cases 6 and 8. Obviously, the simpler non-switching priority setting would be attractive if the difference between its solution value (e.g., the number of mortalities) and the solution value of Problem (P3) is small.

To quantify the advantage of optimal over greedy (non-switching) policies, we sought parameter values that lead to the largest difference. (We restricted attention to parameter ranges that are practically realistic.) The full details are presented in the Online Supplement (noted as Cases 6 and 8—analysis results). Based on our numerical analysis, we expect that when the greedy solution prioritizes the same station that the optimal solution prioritizes first, the incremental cost of using the greedy policy will be very small—less than 0.1% for the worst case. When the greedy solution prioritizes a station that is different from what the optimal solution prioritizes first, then the cost of using the greedy solution may be higher (e.g., 10%) in the worst case. For these cases, which can be identified in advance by comparing the characteristics of Cases 6 and 8 (e.g., for Case 6 priority is first given to Station 2) with the corresponding greedy solution, our recommendation is to solve Problem (P3) and identify explicitly the priority switch time.

The suggested allocation policies to Problem (P1) are listed in Table 2; similar to above, they are classified by the relationships between $\mu_1(1-p_{12})$, μ_2, θ_1, and θ_2.

The entries in the table indicate which station enjoys the higher priority: this high-priority station enjoys all of the surgeons it needs, to the extent possible, while the other station gets the rest, if any. Table 2 covers all of our nine cases, according to the relations between $\mu_1(1-p_{12})$, μ_2, θ_1, and θ_2. It can be used to determine an allocation policy for environments that resemble our model of two serial stations, as well as for other possible types of MCEs, such as road or railroad accidents, radiation or chemical materials leakage, and terrorist bombings. Although it seems difficult to match an MCE type to its optimal allocation policy, the expert trauma doctors with whom we consulted suggested that most accidents and terrorist bombing events could be classified as Case 5. In such events, it is expected that the mortality rate at Station 1 will be higher than in Station 2; the average time of life-saving treatment is estimated at 30 minutes, and about 25% of the casualties will need an operation that lasts about 100 minutes on average. Under such conditions (or similar), Table 2 suggests that an optimal policy would be to prioritize Station 1 throughout the event. It was reassuring to learn from experienced trauma doctors, who had not been exposed to Table 2, that given an additional surgeon during such an event, they would intuitively allocate that surgeon to Station 1.

5. An analysis of two MCE scenarios

In this section we demonstrate the application of Table 2 results to reference MCE scenarios. To be as realistic as possible, we used two scenarios that are based on actual terror attacks. Specifically, we analyzed the results with the medical overseer of the events and found that our model provides logical and coherent management policies.

In both scenarios, the ambulances that were sent to transport the Immediates to the hospital had to return to the MCE scene for remaining casualties. One of the events occurred relatively far from the hospital and the other was close by, which gave rise to differing demand patterns for medical treatment (two waves of arrivals). For the more distant event, the waves of arrivals are twice the life-saving treatment duration from each other; for the close event it takes only half life-saving treatment time from the last first-wave arrival until the first second-wave arrival. The two scenarios (1 and 2) have the following parameters: $\mu_1 = 1/30$, $\mu_2 = 1/90$, $\theta_2 = 1/300$, $\theta_2 = 1/900$, $p_{12} = 0.33$, $N = 10$.

Table 2 indicates that, in both events, the optimal policy is to prioritize life-saving treatments over operations. The optimal policy for the more distant event is presented in Fig. 4(a): upon the arrival of the first wave of casualties, all surgeons are allocated to life-saving treatments; then in anticipation of the second wave, some surgeons continue with their patients to surgery but, within an hour or so, all are again reallocated to life-saving treatments, and so on. In contrast, for the event that occurred near the hospital, the surgeons are allocated to the life-saving treatment policy and move gradually into the operating rooms only when the number of Station 1 casualties gets lower than N. The information gained from allocating surgeons to treatment stations (life-saving versus operating) provides recommendations for management policies that can be used to prepare for similar MCEs. Furthermore, these insights also help resolve clinical/operational tradeoffs (e.g., will a surgeon who performed a life-saving treatment on a casualty also perform the corresponding operation?).

6. Real-time MCE management

In the previous sections we developed a model for planning preparedness for MCEs. It uses reference scenarios (e.g., known casualties' arrival rates) and finds optimal resource allocation policies. The decision makers can draw insights and prepare their staff according to corresponding guidelines. If, for example, according to the relevant reference scenarios, priority should be given to Station 2, then this should be exercised and backed up by appropriate routines. An example of a routine that gives priority to Station 2 and also has clinical advantages is one that guides surgeons, who performed life-saving treatment on a casualty that turns out to need an operation, to continue this casualty's treatment and perform the operation.

In this section we turn to *real-time* decisions. Specifically, we extend the approach that was developed in the previous sections to support real-time allocation decisions in MCEs. In order to ultimately make wise decisions, the common practice in crisis events, such as an MCE, is to perform periodic assessments to review the current status and future forecasts. The information about forecasted arrivals, current load of casualties, and available resources is constantly updated and verified. Standard data analyses of the ongoing event can reveal that parameters

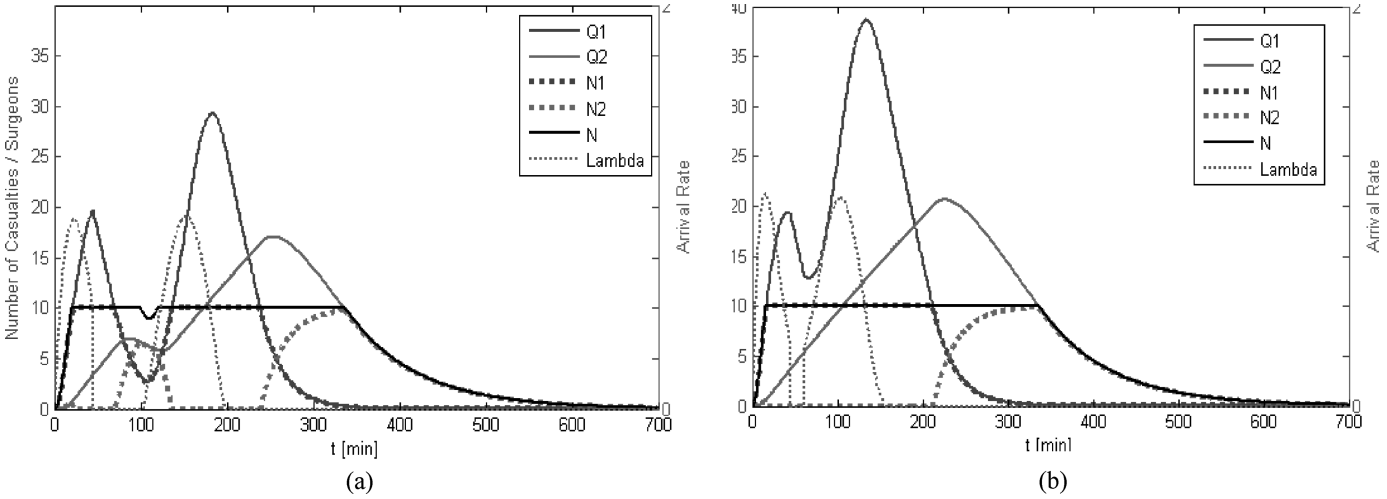

Fig. 4. A dynamic allocation of surgeons to two treatment stations, life-saving followed by operating, so as to minimize mortality during an MCE. (a) This plot represents an event that took place far from the hospital; hence, the arrival waves are 60 minutes apart, and (b) represent an event at closer proximity where the arrival waves are 15 minutes apart. N1 and N2 are the optimal surgeon allocations to Stations 1 and 2. Lambda, Q1, Q2 are casualties' arrival rates and the number of casualties at Stations 1 and 2; N is the total number of allocated surgeons and the maximal number of surgeons is limited to 10 ($N \leq 10$).

such as the arrival rates, actual mortality rates, and treatment times are different than assumed in the reference scenarios (e.g., due to the nature of injuries or misinformation), which may change the resource allocations and sometimes the priorities between the treatment stations. It is important in these cases to take an approach that bases decision making on the most current information. We adopt such an approach. Based on current estimates of the situation, we take a rolling horizon approach that solves the optimization problem (P5) to optimality, at each status assessment, to find the optimal resource allocations from that point onwards. These allocations, as well as the station priorities, may turn out different from those that resulted from the planning models. In this section, we assume that allocations must follow periodic situation assessments that take place every S minutes, where S is the minimal time window in which allocations can be practically changed.

To recapitulate, the main two differences between the planning models introduced earlier and the one presented in the sequel are: first, for real-time management, we use the most updated information about the event, which is constantly being updated and may be different from the information used for the planning models; and second, unlike prior models, we allow the resource allocations to change only after status assessments.

In the formulation of Problem (P2) and Proposition 1, we introduced constraints that force the number of surgeons to be smaller or equal to the number of casualties at a station. Clearly, these constraints are not feasible under the new assumption of periodic allocations. We thus develop a linear formulation, Problem (P5), which includes constraints to ensure that allocations are set periodically and enables a solution to be reached by means of commercial software packages. The details of developing the formulation are presented as Proposition 6 in the Online Supplement.

(P5):

$$\min_{N_1(\cdot),N_2(\cdot)} \sum_{t=0}^{T} [\theta_1 Q_1(t) + \theta_2 Q_2(t)],$$

such that for $t = 0, 1, \ldots, T-1$:

$$Q_1(t+1) = Q_1(t) + \lambda_1(t) - \mu_1 Z_1(t) - \theta_1 Q_1(t),$$

$$Q_2(t+1) = Q_2(t) + p_{12} \times \mu_1 Z_1(t) - \mu_2 Z_2(t) - \theta_2 Q_2(t),$$

$$N_1(t) + N_2(t) \le N,$$

$$Z_i(t) \le N_i(t),\ Z_i(t) \le Q_i(t) \qquad i = 1, 2,$$

$$Z_i(t) \ge 0, N_i(t) \ge 0,\ Q_i(t) \ge,\ Q_i(0) = 0 \quad i = 1, 2,$$

and

$$N_i(u) = N_i(u+1) = \ldots = N_i(u+S-1) \qquad i = 1, 2; u = 0, S, 2S \ldots \lfloor T/S \rfloor S.$$

Variables $Z_i(\cdot)$ are added to formulate the problem as a linear one. Formulation (P5) is updated as time progresses: for example, at the first time assessment the original constraints $Q_1(0) = Q_2(0) = 0$ are replaced by $Q_1(S) = Q_{1,s}$ and $Q_2(S) = Q_{2,s}$. The latter values, $Q_{1,s}$ and $Q_{2,s}$, are determined according to the actual number of people observed at each station at time S, and so the problem is solved from time S onwards (T can be updated and should be long enough to ensure that the last casualty is treated) using the most updated information regarding casualty arrivals, mortality rates, treatment rates, the total number of surgeons, etc.

To make things concrete, we provide two illustrative examples. Their parameters were fit to the outlines of a terror attack that happened in Israel in 2003; parameters such as mortality rates and treatment times were estimated by the event's manager (an M.D.). We manipulated the original arrival rate of the event and assumed that it is a quadratic function to illustrate an event that lasts for several hours. The specific details are as follows:

$$\mu_1 = 1/30,\ \mu_2 = 1/100,\ \theta_1 = 1/300,\ \theta_2 = 1/900,$$
$$p_{12} = 0.25,\ N = 10,$$
$$\lambda(t) = -1 \times 10^{-5} t^2 + 0.0044t,\ 0 \le t \le 440.$$

A status assessment takes place every hour ($S = 60$) in which new information is revealed and new allocations can be made. For the first example, our reconstruction of the status assessments assumes that in the first three assessments the situation appears worse than originally thought, and the forecast for the arrival rate is 10% higher; from that point forward the arrival forecasts do not deviate anymore from the original forecasts. This example fits a common situation whereby in the early stages of an event there is ample uncertainty and confusion as to how it will develop; after some time, however (e.g., after all the casualties have been triaged at the scene, waiting for transport to the hospital), the forecasts prove to be somewhat accurate. Figure 5 describes subsequent solutions of Problem (P5) at four time points ($t = 0$, 60, 120, and 180) and the resulting resource allocation policies. The $t = 0$ solution can be considered as the reference scenario solution and the next solutions at each of the following status assessments take into consideration updated information and allocate to Station 1 more resources for longer time periods. The event's manager can prepare for these changes in the original reference scenario solutions. For this scenario, it is preferable to apply the real-time management solutions over the initial allocation that was found at $t = 0$. The benefit amounts to 24%, whereby reducing the mortality of the Immediates from 35.5 to 27%.

The second example demonstrates a switch in priorities during the MCE due to a change in the event's parameters. Merin *et al*. (2010), who operated a field hospital after the massive earthquake in Haiti, provide evidence of changes in the types of injuries and mortality rates between casualties who arrive first and those who arrive later. For the sake of our example, we assume that during the fourth situation

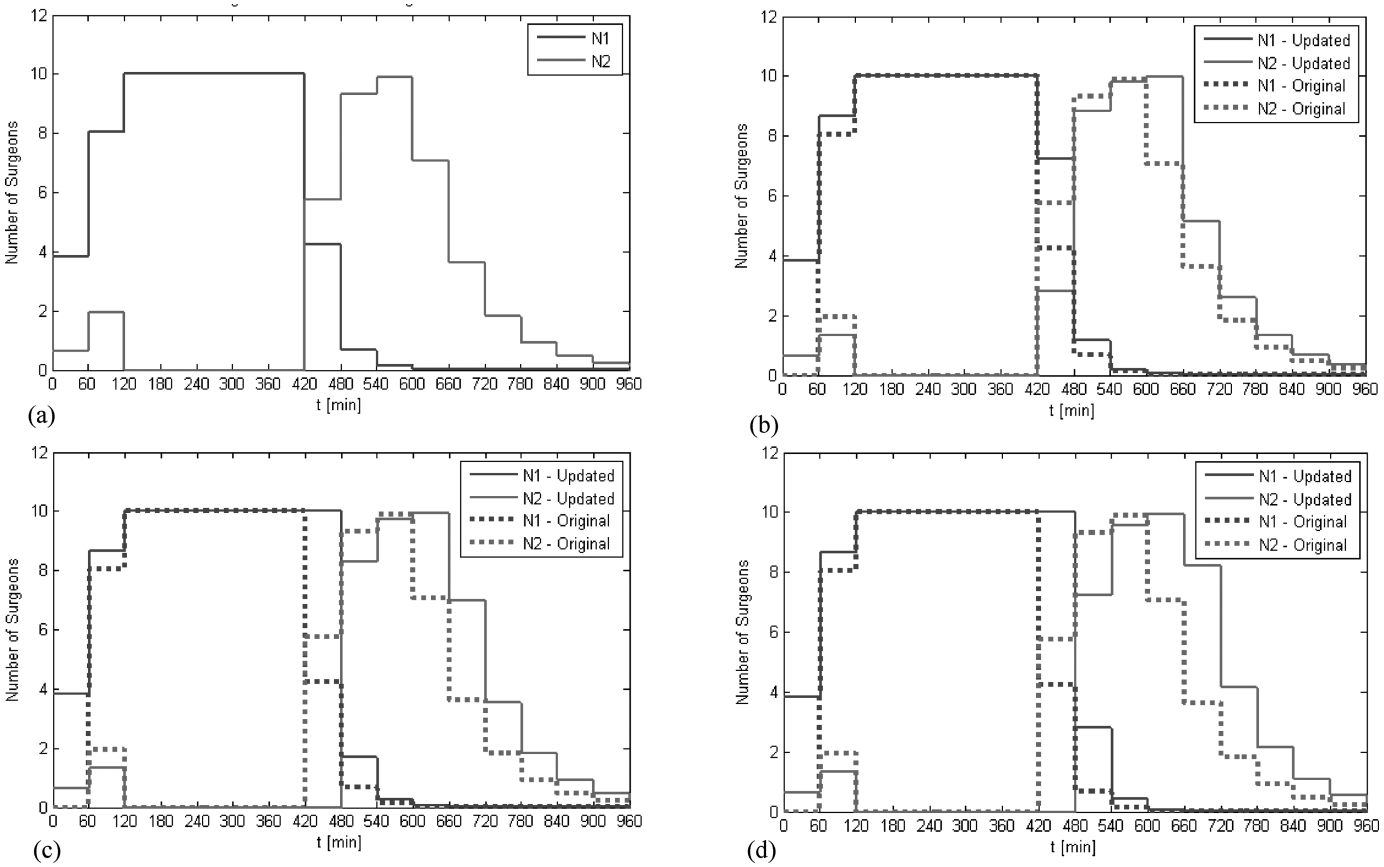

Fig. 5. Optimal resource allocation solutions for different time points 0, 60, 120, 180, corresponding to (a), (b), (c), and (d), respectively. The dotted lines in (b), (c), and (d) represent the resource allocations that were found at $t = 0$ and the solid lines represent the optimal allocations based on updated information.

assessment, 240 minutes after the start of the event, data evaluations revealed that mortality rates are equal at both treatment stations and that more casualties are in need of operations: $\theta_1 = \theta_2 = 1/300$ and $p_{12} = 0.58$. Here, the resource allocation priority that was given so far to Station 1 switches to Station 2. Figure 6 illustrates this situation: the optimal policy from 240 minutes onwards allocates to Station 2 all the resources it needs and the rest go to Station 1. The switch in priorities reduced the mortality by 10.7% compared with the original policy that prioritizes resources to Station 1. This example demonstrates the importance of the real-time management model that alerts the decision maker to possible changes in the allocation policy.

The real-time model considers changes in the environment and informs the decision maker about recommended resource allocation policies. It is adapted to the common practice in crisis events, which calls for periodic status assessments for structured decision making. As is usual for models that support management decisions, here also there are many factors that affect actual decisions, which are not accommodated by our model—for example, capacity utilization at the different stations, waiting times, availability of other resource types, and the specific medical situation of individual casualties.

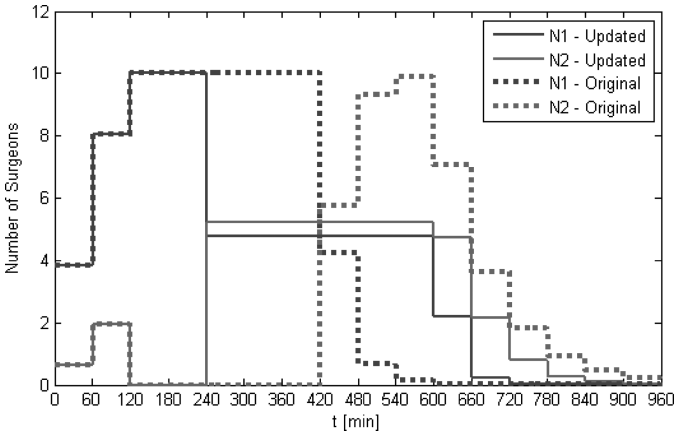

Fig. 6. Optimal resource allocation solutions for time point 0 (dotted line) and 240 (solid line).

7. Summary

Traditional MCE models are based on simulation experiments. Although simulation is a valid modeling tool, it only provides limited structural insights into optimal management policies. We have developed a problem formulation based on fluid models for MCEs. We discretize time and formulate the problem as a LP that enables finding resource allocations for different scenarios as well as structural insights and policies.

We model two stations where surgeons provide medical treatment: a life-saving treatment at the first station and an operation at the second. Based on the relative values of service rates, routing probabilities, and mortality rates, our analysis has identified two types of optimal allocation policies that can be used. One is a static policy that prioritizes one station over the other, and the second is a mixed policy in which priority is first given to one station and at some future time switches to the second.

When a mixed policy is optimal there is no closed-form analytical solution for the priority switch time, but this is easily identified by solving a linear optimization problem. Our proposed policies can be easily applied to prepare an emergency plan for reference scenarios. The resulting allocations represent guidelines that could be used by a manager either in a real-time context or as guidelines on how to prioritize the resources. We developed a rolling horizon approach for the real-time management of MCEs. This approach, which is based on our fluid modeling framework and on management practices for emergency situations, exploits new information about the MCE to create optimal resource allocations. The LP formulation ensures fast solution times—thus, it is practical for use in real-time settings. In MCEs when the uncertainty is high regarding the number of arriving casualties, their treatment rates and other parameters that are changing over time, the real-time approach is in fact essential.

Our model formulation is consistent with the tactical use of fluid models to approximate transient and loaded service networks. We note that our stylized model serves as a proof of concept under our specific assumptions and limitations; for example, constant mortality rates at each of the stations and the surgeons as the only limiting resource. To the best of our knowledge we are introducing mortality rates at a two-station setting for the first time. However, it is important to note that the model does not discriminate between individual casualties based on their specific parameters. Moreover, real-time allocation decisions associated with a specific casualty could be affected by a variety of factors that are not considered in our model; e.g., capacity utilization at the different stations, waiting times, availability of other resources, individual casualties' medical condition, etc. Therefore, other and possibly more refined models can be developed via our approach, sometimes inevitably at the cost of tractable analytical solutions.

Our analysis of numerical examples (based on data extracted from the literature and expert opinions) indicates that our approach can lead to optimal or near-optimal solutions for minimizing the mortality of casualties.

The stylized model introduced in this article has led to the development of structural results and also has provided managerial insights into allocation policies. Its primary message concerns guidelines and quantification of the value of dynamic resource allocation for the management of MCEs. The results presented here provide a starting point for future research that could be based on expanding our modeling scope and testing our policy implications in real-world situations. Analytical extensions can focus on finding the optimal time in which there are priority changes and on the development of bounds on the differences between greedy and optimal policies. Another natural extension is to enhance the model to include other types of MCEs, such as non-conventional MCEs (biological, chemical, nuclear and radiation), where there are different medical processes and resources. Finally, we note that we expect our insights for the optimal non-switching policies to hold for infinite horizon, heavy traffic, two-stage tandem systems, with abandonments and flexible servers. Therefore, this research provides strong motivation for future research on using fluid models to support control of management policies for transient queueing networks in heavy traffic.

Acknowledgements

The authors would like to thank Dr. Shlomi Israelit, present Director, and Dr. Moshe Michaelson, a former Director, of the ED at our partner Rambam Hospital, in Haifa, Israel. We gained a lot from their expertise in MCE management and from many fruitful discussions that helped set up, perform, and improve the present research. We gratefully acknowledge the contribution of the editorial team that made this a better research work.

Funding

The work of A.M. has been partially supported by BSF grants 2005175 and 2008480, ISF grant 1357/08, and the Technion funds for promotion of research and sponsored research. Some of the research was funded by and carried out while A.M. was visiting the Statistics and Applied Mathematical Sciences Institute (SAMSI) of the NSF; the Department of Statistics and Operations Research (STOR), the University of North Carolina at Chapel Hill; the Department of Information, Operations and Management Sciences (IOMS), Leonard N. Stern School of Business, New York University; and the Department of Statistics, The Wharton School, University of Pennsylvania. The wonderful hospitality of these institutions is gratefully acknowledged and truly appreciated.

Supplemental Material

Supplemental data for this article can be accessed on the publisher's website.

References

Ahn, H., Duenyas, I. and Lewis, M.E. (2002) Optimal control of a two-stage tandem queueing system with flexible servers. *Probability in the Engineering and Informational Sciences*, **16**, 453–469.

Altay, N. and Green, W.G. (2007) OR/MS research in disaster operations management. *European Journal on Operational Research*, **175**, 475–493.

Andradottir, S. and Ayhan, H. (2005) Throughput maximization for tandem lines with two stations and flexible servers. *Operations Research*, **53**(3), 516–531.

Argon, N.T., Ziya, S. and Righter, R. (2008) Scheduling impatient jobs in a clearing system within sights on patient triage in mass casualty incidents. *Probability in the Engineering and Informational Sciences*, **22**(3), 301–332.

Atar, R., Mandelbaum, A. and Reiman, M.I. (2004) Scheduling a multi class queue with many exponential servers: asymptotic optimality in heavy traffic. *The Annals of Applied Probability*, **14**(3), 1084–1134.

Aylwin, C.J., Konig, T.C., Brennan, N.W., Shirley, P.J., Davies, G., Walsh, M.S. and Brohi, K. (2006) Reduction in critical mortality in urban mass casualty incidents: analysis of triage, surge, and resource use after the London bombings on July 7, 2005. *Lancet*, **368**, 2219–2225.

BBC News. (2012) Argentine train crash: brake warning denied. Available at http://www.bbc.co.uk/news/world-latin-america-17174635 (accessed February 27, 2012).

Einav, S., Aharonson-Daniel, L., Weissman, C., Freund, H.R., Peleg, K. and Israel Trauma Group. (2006) In-hospital resource utilization during multiple casualty incidents. *Annals of Surgery*, **243**(4), 533–540.

Green, L.V., Kolesar, P.J. and Whitt, W. (2007) Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, **16**(1), 13–39.

Hick, J.L., Hanfling, D., Burstein, J.L., DeAtley, C., Barbisch, D., Bogdan, G.M. and Cantrill, S. (2004) Health care facility and community strategies for patient care surge capacity. *Annals of Emergency Medicine*, **44**(3), 253–261.

Hirshberg, A., Frykberg, E.R., Mattox, K.L., and Stein, M. (2010) Triage and trauma workload in mass casualty: a computer model. *Journal of Trauma-Injury Infection & Critical Care*, **69**, 1074–1082.

Hirshberg, A., Holcomb, J.B. and Mattox, K.L. (2001) Hospital trauma care in multiple-casualty incidents: a critical view. *Annals of Emergency Medicine*, **37**, 647–652.

Hirshberg, A., Scott, B.G., Granchi, T., Wall Jr., M.J., Mattox, K.L. and Stein, M. (2005) How does casualty load affect trauma care in urban bombing incidents? A quantitative analysis. *Journal of Trauma*, **58**, 685–695.

Hirshberg, A., Stein, M. and Walden, R. (1999) Surgical resource utilization in urban terrorist bombing: a computer simulation. *Journal of Trauma*, **47**, 545–550.

Hughes, M.A. (1991) A selected annotated bibliography of social science research on planning for and responding to hazardous material disasters. *Journal of Hazardous Materials*, **27**, 91–109.

Jacobson, E.U., Argon, N.T. and Ziya, S. (2012) Priority assignment in emergency response. *Operations Research*, **60**(4), 813–832.

Kellerer, H., Pferschy, U. and Pisinger, D. (2004) *Knapsack Problems*. Springer-Verlag, Berlin, Germany.

Kosashvili, Y., Aharonson-Daniel, L., Peleg, K., Horowitz, A., Laor, D. and Blumenfield, A. (2009) Israeli hospital preparedness for terrorism-related multiple casualty incidents: can the surge capacity and injury severity distribution be better predicted? *Injury*, **40**(7), 727–731.

Kreiss, Y., Merin, O., Peleg, K., Levy, G., Vinker, S., Sagi, R., *et al.* (2010) Early disaster response in Haiti: the Israeli field hospital experience. *Annals of Internal Medicine*, **153**(1), 45–48.

Lerner, E.B., Schwartz, R.B., Coule, P.L., Weinstein, E.S., Cone, D.C., Hunt, R.C., *et al.* (2008) Mass casualty triage: an evaluation of the data and development of a proposed national guideline. *Disaster Medicine and Public Health Preparedness*, **2**(1), S25–S34.

Mandelbaum, A., Massey, W.A., Reiman, M.I. and Rider, R. (1999) Time-varying multiserver queues with abandonments and retrials, in *Proceedings of the 16th International Teletra Congress*, Key, P. and Smith, D. (eds), pp. 355–364.

Mandelbaum, A., Massey, W.A., Reiman, M.I. and Stolyar, A. (1999) Waiting time asymptotics for time varying multiserver queues with abandonment and retrials in *Proceedings of the Annual Allerton Conference on Communication Control and Computing*, pp. 1095–1104.

Mehta, S. (2006) Disaster and mass casualty management in a hospital: how well are we prepared? *Journal of Postgraduate Medicine*, **52**(2), 89–90.

Merin, O., Ash, N., Levy, G., Schwaber, M.J. and Kreiss, Y. (2010) The Israeli field hospital in Haiti—ethical dilemmas in early disaster response. *New England Journal of Medicine*, **362**, e38(1)–e38(3).

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

Oliver, R.M. and Samuel, A.H. (1962) Reducing letter delays in post offices. *Operations Research*, **10**, 839–892.

Pang, G. and Whitt, W. (2012) The impact of dependent service times on large-scale service systems. *Manufacturing & Service Operations Management*, **14**(2), 262–278.

Paul, J.A., George, S.K., Yi, P. and Lin, L. (2006) Transient modelling in simulation of hospital operations for emergency response. *Prehospital and Disaster Medicine*, **21**(4), 223–236.

Sinreich, D. and Marmor, Y. (2004) A simple and intuitive simulation tool for analyzing emergency department operations, in *Proceedings of the 2004 Winter Simulation Conference*, IEEE Publications, Piscataway, NJ, pp. 1994–2002.

Vandergraft, J.M. (1983) A fluid flow model of networks of queues. *Management Science*, **29**(10), 1198–1208.

Waeckerle, J.F. (1991) Disaster planning and response. *New England Journal of Medicine*, **324**, 815–821.

Whitt, W. (2005) Two fluid approximations for multi-sever queues with abandonments. *Operations Research Letters*, **33**, 363–372.

Whitt, W. (2006) Fluid models for multiserver queues with abandonments. *Operations Research*, **54**(1), 37–54.

Yom-Tov, G. (2010) Queues in hospitals: queueing networks with reentering customers in the QED regime. Ph.D. Thesis, IE&M, Technion, Israel.

Biographies

Izack Cohen is a visiting lecturer in the Industrial Engineering and Management Faculty at the Technion. He received his B.Sc. in Chemical Engineering (1991), his M.Sc. in Materials Engineering (1997), and his Ph.D. in Industrial Engineering and Management (2004). He has managed various technological, logistics, and information systems organizations and technological projects. His research interests include project management, supply chain management, service engineering for healthcare, and innovative maintenance methods.

Avi Mandelbaum is the Benjamin and Florence Free Chair Professor of Operations Research, Statistics and Service Engineering, at the Faculty of Industrial Engineering (IE&M) at the Technion, Israel. He has a B.Sc. in

Mathematics and Computer Science and an M.A. in Statistics, both from Tel-Aviv University. His Ph.D. is in Operations Research, from Cornell University. After graduation he joined the Graduate School of Business at Stanford University. He left the United States, in 1991 to assume a position at IE&M. This faculty position, as well as his current research and application interests, constitutes a convex hull of all the areas in which he had been educated. He founded and has been the academic director of the Technion IE&M SEELab (SEE = Service Enterprise Engineering). The SEELab is collecting and maintaining data from large service operations, which it then prepares to support research and teaching.

Noa Zychlinski has a B.Sc. (*Cum laude*, 2005) and an M.Sc. (*Cum laude*, 2012) in Industrial Engineering and Management from the Technion, Israel. After graduating from her B.Sc., she worked as a project manager and a team leader in an IT company that develops and implements organizational information systems. Her research interests include service systems, with focus on healthcare applications.