Ann Oper Res (2016) 236:177–196
DOI 10.1007/s10479-014-1716-1

A medical resource allocation model for serving emergency victims with deteriorating health conditions

Yisha Xiang · Jun Zhuang

Published online: 14 September 2014

Abstract Large-scale disasters typically result in a shortage of essential medical resources, and thus it is critical to optimize resource allocation to improve the quality of the relief operations. One important factor that has been largely neglected when optimizing the available medical resources is the deterioration of victims' health condition in the aftermath of a disaster; e.g., a victim's health condition could deteriorate from mild to severe if not treated promptly. In this paper, we first present a novel queueing network to model this deterioration in health conditions. Second, we provide both analytical solutions and numerical illustrations for this queueing network. Finally, we formulate two resource allocation models in order to minimize the total expected death rate and total waiting time, respectively. Numerical examples are provided to illustrate the properties of optimal policies.

Keywords Deterioration · Disaster · Emergency relief · Queueing network · Health care

1 Introduction

Man-made and natural disasters cause a large number of casualties, especially in highly inhabited areas (Zhuang and Bier 2007). The Indian Ocean tsunami in 2004 killed more than

The work of the first author was supported by Chinese Ministry of Education under Grant 11YJC630228, and Natural Science Foundation of Guangdong under Grant S2011040002092. The work of the second author was partially supported by the United States Department of Homeland Security (DHS) through the National Center for Risk and Economic Analysis of Terrorism Events (CREATE) under award number 2010-ST-061-RE0001, and by the United States National Science Foundation under award numbers 1200899 and 1334930. However, any opinions, findings, and conclusions or recommendations in this document are those of the authors and do not necessarily reflect views of the DHS, CREATE, or NSF.

Y. Xiang
The School of Business, Sun Yat-sen University, Guangzhou, Guangdong, China

J. Zhuang (✉)
Department of Industrial and Systems Engineering, University at Buffalo, Buffalo, NY, USA
e-mail: jzhuang@buffalo.edu

225,000 people and relocated millions more in countries spread around the Ocean's rim from Kenya to Indonesia (Altay and Green 2006). Hurricane Katrina made landfall along the Gulf Coast on Aug 25, 2005. At least 1,836 people died in the hurricane and in the subsequent floods, and the total damage from Katrina is estimated at $81 billion (2005 U.S. dollars). The Great Sichuan Earthquake in 2008 was a deadly earthquake that measured at 8.0 magnitude, killed about 70,000 people, and left more than 18,000 missing. Recently, the 2011 Japanese earthquake/tsunami/nuclear crisis caused enormous casualties and economic losses. Statistics show that more than 255 million people are affected annually by disasters. Providing medical care to the victims is a daunting task (Kahn et al. 2009) and requires partnership from public and private sectors (Hausken and Zhuang 2013). In addition, large-scale catastrophic events often result in a scarcity of essential medical resources, including funding (Zhuang et al. 2014), supplies, equipment, facilities, and personnel. The allocation of limited resources and essential services becomes critical in ensuring that all affected individuals are provided with the best possible opportunities for survival while sustaining overall societal function and stability (Bostick et al. 2008).

Despite the great need to better assign and schedule available medical resources to minimize the loss of life and maximize the efficiency of the rescue operations, a review of the literature shows that there is a primary focus on logistics management in the aftermath of disasters (Haghani and Oh 1996; Barbarosolu et al. 2002; Ozdamar et al. 2004; Yi and Kumar 2007; Sheu 2007, 2010; Coles et al. 2012), whereas medical resource allocation problems have received much less attention (Fiedrich et al. 2000; Gong and Batta 2007). Many state-of-the-art tools for supporting emergency health management (e.g., HAZUS) usually focus on information systems, and do not provide informative decision support (Fiedrich et al. 2000). These systems locate and classify the available resources, but are usually not able to provide optimal resource allocation plans.

In practice, triage is one commonly used tool in disaster and emergency medicine for large-scale catastrophic events. In the 1980's, one of the first civilian triage systems was developed in U.S., known as simple triage and rapid treatment (START) (Super et al. 1994). START was rapidly adopted across the United States and in some international settings as well. It proved useful in prioritizing the transportation of the most critical patients, but it ignores the availability of resources and the deterioration of patients' conditions. Other triage systems include the Triage Sieve, the Care Flight Triage, and the Sacco Treatment Method (STM) (Jenkins et al. 2008). Sacco et al. (2005) are the first ones that explicitly consider resource constraints in triage management, which mathematically formulates a resource-constrained triage problem with the goal of maximizing the expected number of survivors, subject to constraints on the timing and availability of transportation and treatment resources. The Delphi technique is used in Sacco et al. (2005) to estimate the deterioration of the victims' health. The limitations of the model proposed in Sacco et al. (2005) are that the number of patients to be evacuated is deterministic, patients who die on their way to the hospital are not considered, and no decision support for resource allocation after the victims are admitted to health care facilities is considered. Hick et al. (2009) propose a taxonomy within surge capacity of conventional, contingency and crisis capacities, and proposed adaptive strategies for staff and supply challenges. Surge capacity generally refers to the ability to manage a sudden, unexpected increase in patient volume that would otherwise severely challenge or exceed the present capacity of the facility. Hick et al. (2009) examine surge capacity primarily in the context of responses within the hospital's physical structure or on the grounds that are managed and staffed by the hospital. Sinuff et al. (2004) examine the impact of rationing intensive care unit beds, and suggest that patients who are perceived to not benefit from critical care are more often refused intensive care unit admission. Cookson and Dolan (2000)

and Dolan and Cookson (2000) qualitatively study the different substantive principles of justice for making health care priority-setting decisions, such as need, equity and fairness principles. However, quantitative tools to help optimize resource allocation have yet to be developed.

The problem of treating victims at different severity levels has been mainly modeled using queueing theories in the field of operations research. Gong and Batta (2006) consider a dynamic disaster environment in which thousands of casualties need to be treated. They develop a two-priority, single-server queueing model, and propose a queue-length cutoff method to minimize the weighted average number of patients in the system. Argon and Ziya (2009) are the first to explicitly consider imperfect customer information on the identities of customers and priority assignment decisions. In particular, Argon et al. (2009) study a network of parallel service stations, each modeled as a single server queue; each station serves its own dedicated customers as well as generic customers, and the model is concerned with the dynamic routing of incoming customers to one of several parallel service stations.

In this study, we consider a different decision approach to medical resource allocation problems in the aftermath of a natural disaster. More specifically, the decision variables here are service rates dedicated to each queue. Similar problems dealing with control of service rates can be found in Cabrill (1974), Weber and Stidham (1987), and George and Harrison (2001). Glazebrook et al. (2004) and Li and Glazebrook (2010) discuss the optimal resource allocation of service to impatient tasks. Our model bears close resemblance to the one proposed by Gong and Batta (2006) in classifying each victim into either a high-priority class (life threatening) or a low-priority class (not life threatening), and assuming poisson arrival process and exponential service times. Our model differentiates from the aforementioned models in that:

- Deterioration of victims' health condition is explicitly modeled, which has not been adequately addressed in the literature; one exception is Li and Glazebrook (2010).
- Decision variables are the amount of service rates instead of traditional queue-length cutoff or server-cutoff.

This paper contributes to the literature by presenting a novel resource allocation model for a patient queueing network with deteriorating health conditions. This would provide new "policy analytics" modeling tools and insights for government departments, hospitals, and other health care agencies in dealing with patient flows. The remainder of this paper is organized as follows. Section 2 formulates the model. Section 3 presents two optimization models, minimizing the expected number of fatalities and the weighted sum of system time, respectively, and conducts sensitivity analyses. Section 4 presents the conclusions and possible directions for future research. Appendix 1 provides the results comparing the decomposition versus the numerical methods. Appendix 2 provides evidence supporting the optimality of the algorithm used in this paper.

2 Model formulation

After a disaster occurs, hundreds or thousands of casualties need to be treated. The casualties in such a disaster setting are usually placed into four severity levels. For example, START algorithm assigns disaster victims into green, yellow, red and black triage categories (Kahn et al. 2009). We focus on the middle two types of victims which are of primary concern of medical rescue operations: type 2 (mild) victims require hospitalization but are not considered life threatening, and type 1 (severe) victims require hospitalization and their conditions can

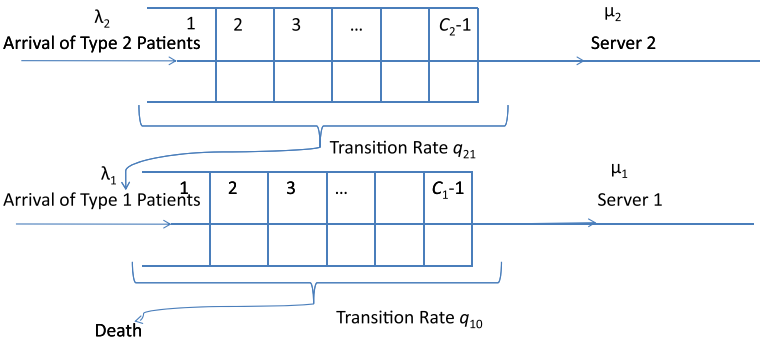

Fig. 1 Illustration of transitions in patient health conditions between queues

become life threatening. For each of the two types of victims, we model the arrival and service process using one single-server queue. Since a multi-server queue can be mapped to a single server queue (Xiong and Altiok 2009), we study the service rates rather than the number of servers for mathematical simplicity. Different victim types require different treatments, so victims are assumed to form separate queues (Argon and Ziya 2009), and do not share servers. It is assumed that the two types of victims arrive according to Poisson processes with rates λ_1 and λ_2, respectively. We acknowledge that the arrivals of victims in many aftermath disaster situations may not strictly follow Poisson processes, and there is uncertainty about victims' arrival rates (Insua et al. 2012). However, for analytical tractability and focusing on the key contribution of this paper, we do not consider heterogeneous arrival processes.

Conditions of patients may deteriorate during the waiting period, and thus cannot be treated by the resources allocated to the queues they initially joined. More specifically, patients with mild conditions (type 2) may get worse if not treated promptly, and become patients with severe conditions (type 1). Similarly, patients with severe conditions may die while waiting for treatment. However, there is no death in patients with less severe conditions since they are not life threatening. We assume that the deterioration process of the patients' conditions can be modeled by continuous time Markov chains. That is, type 2 patients evolve to type 1 with rate q_{21}, type 1 patients evolve to type 0 (death) with rate q_{10}; and there is no direct transition from type 2 to type 0. For each queue, service times for patients are independent and exponentially distributed with rates μ_1 and μ_2, respectively, as illustrated in Fig. 1. We assume finite capacities C_1 and C_2 for type 1 and type 2 queues, respectively. The arrival rates of victims could be estimated by collecting the actual data from historical disasters. The service rates could be difficult to obtain, since it is generally difficult to estimate how long the medical service will last and most post-disaster medical records focus on the number of injuries and deaths instead. However, possible methods to estimate service rate include doctor estimation, analysis of historical durations, adjusting for case complicity, and combination of the above (Macario 2010). Similarly, the deterioration rates from mild conditions to severe conditions could be estimated using historical data, expert opinion, as well as brainstorming methods.

2.1 Notation and model formulation

Throughout this paper, we use the following notation:

- λ_i: Arrival rate for type i patients, $i = 1, 2$.
- R: Total budget available.

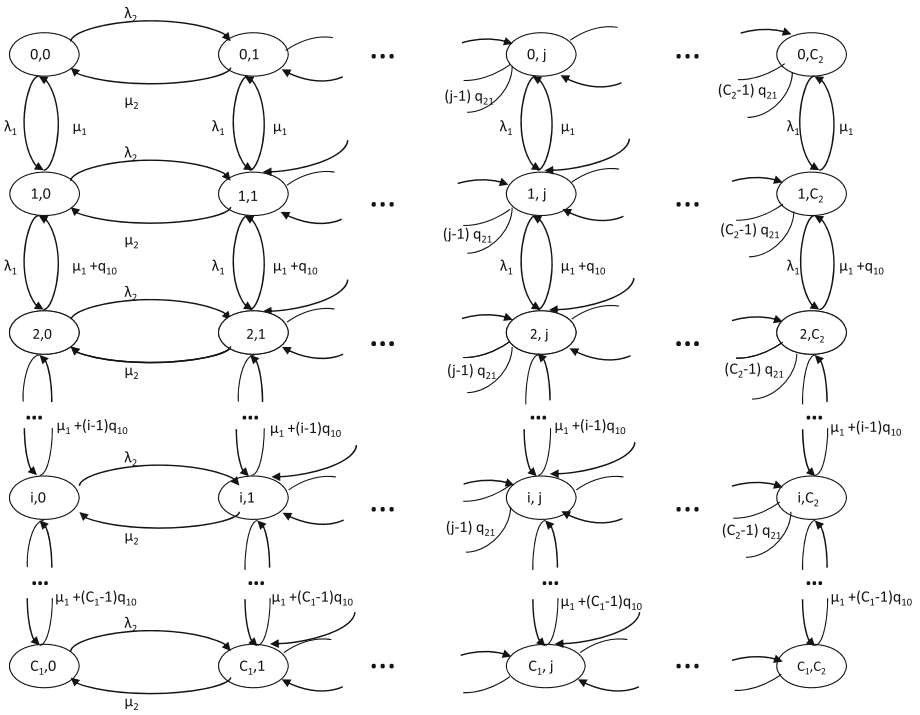

Fig. 2 Transition rate diagram

- π_i: Unit cost of service rate for type i patients, $i = 1, 2, \pi_1 \geq \pi_2$.
- μ_i: Service rate allocated to type i patients, $i = 1, 2$.
- q_{ij}: Deterioration rate from type i to type j, $i = 1, 2$ and $j = i - 1$.
- L_i: Average number of type i patients in the system, $i = 1, 2$.
- W_i: Average system time of type i patients, $i = 1, 2$.
- N_d: Death rate due to delay in treatment, excluding lost patients.
- C_k: Capacity limit for type k queue, $k = 1, 2$.
- N_k: Number of type i patients in the system, $N_k = 0, 1, \ldots, C_k, k = 1, 2$.
- K and $1 - K$: Coefficient/weight assigned to the system time of types 1 and 2 patients.
- $P(i, j)$: Steady-state probabilities that there are i type 1 patients and j type 2 patients in the system.
- $P(N_k = n)$: Steady-state probabilities that there are n type i patients in the system, $n = 0, 1, 2, \ldots, C_k, k = 1, 2$.
- $P_{C_k} = P(N_k = C_k)$: The probability that type k queue at its full capacity, $k = 1, 2$.

Let i and j be the number of type 1 and type 2 patients in the system at time t, respectively. We consider a bivariate process $\{(N_1(t), N_2(t)), N_1(t) = 0, 1, \ldots, C_1, N_2(t) = 0, 1, \ldots, C_2, t \geq 0\}$ with state space $S = \{(i, j), i = 0, 1, \ldots, C_1, j = 0, 1, \ldots, C_2\}$. The service discipline is assumed to be first-come, first-serve (FCFS) for each patient type with no sharing between the two servers. The transition rate diagram is illustrated in Fig. 2.

Under steady-states, we have the following balance equations for all sets of states for $i = 1, \ldots, C_1$, and $j = 1, \ldots, C_2$:

- for $i = 0$, and $j = 0$:

$$(\lambda_1 + \lambda_2)P(0,0) = \mu_1 P(1,0) + \mu_2 P(0,1)$$

- for $i = 0$, and $j = 1, 2, \ldots, C_2 - 1$:

$$(\lambda_1 + \lambda_2 + (j-1)q_{21} + \mu_2)P(0,j) = \lambda_2 P(0,j-1) + \mu_1 P(1,j) + \mu_2 P(0,j+1)$$

- for $i = 0$, and $j = C_2$:

$$(\lambda_1 + (C_2 - 1)q_{21} + \mu_2)P(0,C_2) = \lambda_2 P(0,C_2-1) + \mu_1 P(1,C_2)$$

- for $i = 1, 2, \ldots, C_1 - 1$, and $j = 0$:

$$(\lambda_1 + \lambda_2 + \mu_1 + (i-1)q_{10})P(i,0) = \lambda_1 P(i-1,0) + (\mu_1 + iq_{10})P(i+1,0) + \mu_2 P(i,1)$$

- for $i = 1, 2, \ldots, C_1 - 1$, and $j = 1, 2, \ldots, C_2 - 1$:

$$\begin{aligned}&(\lambda_1 + \lambda_2 + (j-1)q_{21} + \mu_1 + (i-1)q_{10} + \mu_2)P(i,j)\\ &= \lambda_1 P(i-1,j) + \lambda_2 P(i,j-1) + jq_{21}P(i-1,j+1)\\ &\quad + (\mu_1 + iq_{10})P(i+1,j) + \mu_2 P(i,j+1)\end{aligned}$$

- for $i = 1, 2, \ldots, C_1 - 1$, and $j = C_2$:

$$\begin{aligned}&(\lambda_1 + (C_2-1)q_{21} + \mu_1 + (i-1)q_{10} + \mu_2)P(i,C_2) = \lambda_1 P(i-1,C_2)\\ &+ \lambda_2 P(i,C_2-1) + (\mu_1 + iq_{10})P(i+1,C_2)\end{aligned}$$

- for $i = C_1$, and $j = 0$:

$$(\lambda_2 + \mu_1 + (C_1-1)q_{10})P(C_1,0) = \lambda_1 P(C_1-1,0) + \mu_2 P(C_1,1)$$

- for $i = C_1$, and $j = 1, 2, \ldots, C_2 - 1$:

$$\begin{aligned}&(\lambda_2 + \mu_1 + (C_1-1)q_{10} + \mu_2)P(C_1,j) = \lambda_1 P(C_1-1,j)\\ &+ \lambda_2 P(C_1,j-1) + jq_{21}P(C_1-1,j+1) + \mu_2 P(C_1,j+1)\end{aligned}$$

- for $i = C_1$, and $j = C_2$:

$$(\mu_1 + (C_1-1)q_{10} + \mu_2)P(C_1,C_2) = \lambda_1 P(C_1-1,C_2) + \lambda_2 P(C_1,C_2-1)$$

2.2 Decomposition and analysis on type 2 queue

In this subsection, we solve the global balance equations in Sect. 2.1 for two-dimensional steady-state distribution by decomposing queues. Note that since only type 2 patients can evolve to type 1 patients, but not vice versa, the queue of type 2 patients is independent of the queue of type 1 patients. Therefore, we are able to analyze the steady state distribution of the queue of type 2 patients. Then, we analyze the conditional steady-state distribution of the queue of type 1 patients, conditional on each of the possible steady-state number of patients in the type 2 queue. Finally, we study the unconditional steady-state distribution of the type 1 queue by using point-wise stationary approximations in time varying queues.

Note that we acknowledge in principle we could directly solve the equations list in Sect. 2.1; however, the number of equations needed to be directly solved for this 2-D Markov chain would be $C_1 \times C_2$. For example, if capacity isf 1,000 for both queues, we need to solve 1,000,000 equations to get the steady-state probability matrix $P(i, j)$; and then we need to

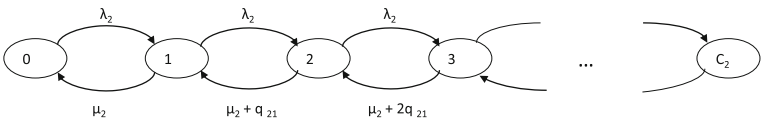

Fig. 3 Type 2 queue transition rate diagram

aggregate those one million probabilities to calculate some system measures such as L and W. Aggregating one million tiny probabilities could generate huge technical computational errors, which may be even larger than our approximation method. On the other hand, solving those huge equation systems multiple times (each time for each step of search during the optimization process) makes the resource allocation optimization almost impossible. By contrast, our proposed decomposition method below provides exact formulas for the solution, independent of the sizes of C_1 or C_2, making the resource allocation optimization possible (see Sect. 3). To verify the approximation method, we numerically solved the two-dimensional balance equations and compared the answers to our approximation. The results show our approximation approach is good, as documented in the Appendix 1.

We begin our decomposition by first analyzing the type 2 queue. Figure 3 shows the transition rate diagram of type 2 patients; and the global balance equations are as follows for $j = 0, 1, \ldots, C_2$:

- for $j = 0$:

$$\lambda_2 P(N_2 = 0) = \mu_2 P(N_2 = 1)$$

- for $j = 1, \ldots, C_2 - 1$:

$$[\lambda_2 + \mu_2 + (j-1)q_{21}]P(N_2 = j) = \lambda_2 P(N_2 = j-1) + (jq_{21} + \mu_2)P(N_2 = j+1)$$

- for $j = C_2$:

$$[\mu_2 + (C_2 - 1)q_{21}]P(N_2 = C_2) = \lambda_2 P(N_2 = C_2 - 1)$$

When $j \geq 1$, probability of j type 2 patients in the system is derived by induction:

$$P(N_2 = j) = \frac{\lambda_2^j}{\prod_{m=1}^{j}[\mu_2 + (m-1)q_{21}]} P(N_2 = 0), \tag{1}$$

According to Eq. (1) and the normalization condition, $\sum_{j=0}^{C_2} P(N_2 = j) = 1$, we have:

$$P(N_2 = 0) = \frac{1}{1 + \sum_{j=1}^{C_2} \frac{\lambda_2^j}{\prod_{m=1}^{j}[\mu_2 + (m-1)q_{21}]}} \tag{2}$$

The expected number of type 2 patients in the system is:

$$L_2 = \sum_{j=1}^{C_2} j P(N_2 = j), \tag{3}$$

And the expected time type 2 patients spend in the system is:

$$W_2 = \frac{L_2}{\lambda_2} = \frac{1}{\lambda_2} \sum_{j=1}^{C_2} j P(N_2 = j). \tag{4}$$

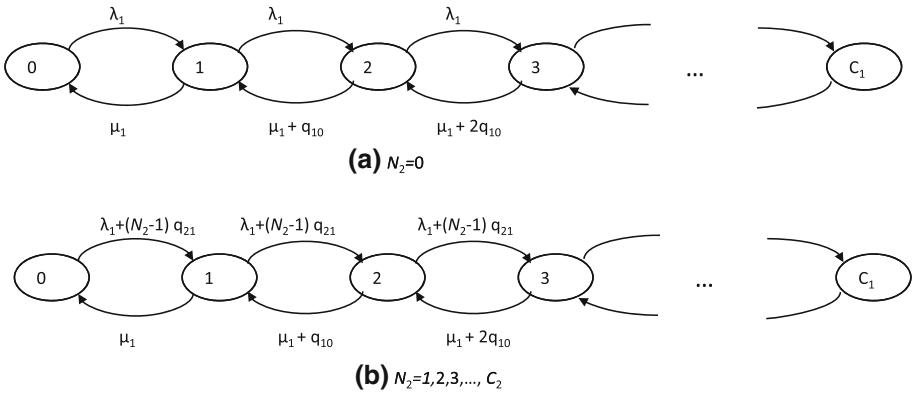

Fig. 4 Type 1 queue transition rate diagram

2.3 Decomposition and analysis on the type 1 queue

We conduct a similar performance analysis on the queue of type 1 patients. Given the number of type 2 patients in the system ($N_2 = j$), the queue of type 1 patients is similar to the type 2 queue. Figure 4 shows the transition rate diagram for the type 1 patients.

Given the number of type 2 patients j in the system, we have the following limiting probabilities:

$$P(N_1 = i|N_2 = j) = \begin{cases} \frac{\lambda_1^i}{\prod_{n=1}^{i}[\mu_1+(n-1)q_{10}]} P(N_1 = 0|N_2 = j) & \text{when } j = 0, 1 \\ \frac{(\lambda_1+(j-1)q_{21})^i}{\prod_{n=1}^{i}[\mu_1+(n-1)q_{10}]} P(N_1 = 0|N_2 = j) & \text{when } j = 2, 3, \ldots, C_2 \end{cases} \tag{5}$$

Since we have $\sum_{i=0}^{C_1} P(N_1 = i|N_2 = j) = 1$, Eq. (5) implies:

$$P(N_1 = 0|N_2 = j) = \begin{cases} \frac{1}{1+\sum_{i=1}^{C_1} \frac{\lambda_1^i}{\prod_{n=1}^{i}(\mu_1+(n-1)q_{10})}} & \text{when } j = 0 \\ \frac{1}{1+\sum_{i=1}^{C_1} \frac{(\lambda_1+(j-1)q_{21})^i}{\prod_{n=1}^{i}(\mu_1+(n-1)q_{10})}} & \text{when } j = 1, 2, 3, \ldots, C_2 \end{cases} \tag{6}$$

The probability that there are i patients in the type 1 system is calculated as follows, using the approximation that we discussed in the beginning of Sect. 2.2:

$$P(N_1 = i) = \sum_{j=0}^{C_2} P(N_1 = i|N_2 = j)P(N_2 = j), \quad \text{for} \quad i = 0, 1, \ldots, C_1. \tag{7}$$

Given the number of type 2 patients in the queue, the average number of type 1 patients in the system $L_1|_{N_2=j}$ is calculated as:

$$L_1|_{N_2=j} = \sum_{i=1}^{C_1} i\, P(N_1 = i|N_2 = j) \tag{8}$$

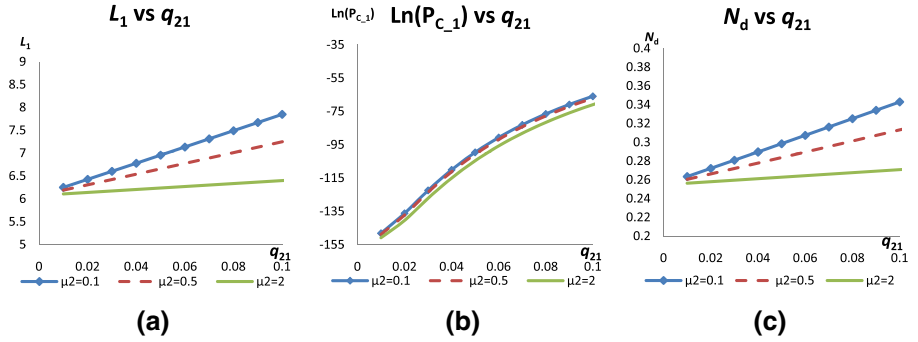

Fig. 5 Effect of transition rate q_{21} on type 1 queue

Therefore, the expected number of type 1 patients in the system (L_1) is calculated as, using the same approximation that we discussed in the beginning of Sect. 2.2:

$$L_1 = \sum_{j=0}^{C_2} P(N_2 = j) L_1|_{N_2=j} = \sum_{j=0}^{C_2} \left[P(N_2 = j) \sum_{i=1}^{C_1} i P(N_1 = i|N_2 = j) \right] \tag{9}$$

Since type 2 patients transition to type 1 patients due to delay in treatment, we need to adjust the arrival rate of the type 1 queue to account for that. Using Little's formula, the expected system time for a type 1 patient in the system, W_1, is calculated as:

$$\begin{aligned} W_1 &= \frac{L_1}{\sum_{j=1}^{C_2} P(N_2 = j)(\lambda_1 + q_{21}(j-1)) + \lambda_1 P(N_2 = 0)} \\ &= \frac{\sum_{j=0}^{C_2} \left[P(N_2 = j) \sum_{i=1}^{C_1} i P(N_1 = i|N_2 = j) \right]}{\sum_{j=1}^{C_2} P(N_2 = j)(\lambda_1 + q_{21}(j-1)) + \lambda_1 P(N_2 = 0)} \end{aligned} \tag{10}$$

The death rate due to lack of treatment is:

$$N_d = \sum_{i=1}^{C_1} P(N_1 = i) q_{10}(i-1), \tag{11}$$

$$= q_{10} \left[L_1 - (1 - P(N_1 = 0)) \right]. \tag{12}$$

Substituting Eq. (7) into Eq. (12), we have:

$$N_d = q_{10} \left[L_1 - 1 + \sum_{j=0}^{C_2} P(N_1 = 0|N_2 = j) P(N_2 = j) \right] \tag{13}$$

Since the type 1 queue is dependent on the type 2 queue, we first analyze the interaction between the queues. We begin our analysis with the effect of q_{21} on the type 1 queue. Suppose $\lambda_1 = 0.5$, $\lambda_2 = 1$, $\mu_1 = 0.5$, $q_{10} = 0.05$, and $C_1 = C_2 = 100$. Three different service rates for the type 2 queue ($\mu_2 = 0.1, \mu_2 = 0.5$, and $\mu_2 = 2$) are arbitrarily selected for comparison. Figure 5a, b, c show that all three performance measures increase in q_{21} and decrease in μ_2. This is because both larger q_{21} and smaller μ_2 would result in more patients transitioning to the type 1 queue, resulting in fewer patients remaining in the type 2 queue.

3 Optimization model and numerical examples

3.1 Optimization models

The main concerns during disaster relief operations are the total death rate and timely medical treatment. We develop two optimization models to minimize the total death rate and weighted total system time, respectively. The first optimization model minimizes the total death rate under a budget constraint. The total death rate includes the summation of the death rate for type 1 patients (N_d), the loss rate of type 2 patients from the queue ($\lambda_2 P_{C_2}$), and the loss rate of type 1 patients from the queue ($\lambda_1 P_{C_1}$).

$$\min_{\mu_1,\mu_2} \quad N_d + \left[\lambda_1 + q_{21}\sum_{j=1}^{C_2} P(N_2 = j)(j-1)\right] P_{C_1} + \lambda_2 P_{C_2} \tag{14}$$
$$\pi_1\mu_1 + \pi_2\mu_2 \le R$$

The minimization problem is subject to budget constraints; thus, equality is always desired. We eliminate the equality budget constraint by linear transformation. Let,

$$\mu_2 = (R - \pi_1\mu_1)/\pi_2 \tag{15}$$

Substituting Eq. (15) into the objective function, we get the following problem, referred to as **P1**:

$$\min_{\mu_1} \quad N_d + \left[\lambda_1 + q_{21}\sum_{j=1}^{C_2} P(N_2 = j)(j-1)\right] P_{C_1} + \lambda_2 P_{C_2} \tag{16}$$
$$0 \le \mu_1 \le R/\pi_1 \tag{17}$$

Secondly, we develop an optimization model to minimize the weighted sum of expected system time, referred to as **P2**:

$$\min_{\mu_1} \quad KW_1 + (1-K)W_2 \tag{18}$$
$$0 \le \mu_1 \le R/\pi_1 \tag{19}$$

Thirdly, we consider an optimization model to minimize the total expected system time, referred to as **P3**:

$$\min_{\mu_1} \quad L_1W_1 + L_2W_2 \tag{20}$$
$$0 \le \mu_1 \le R/\pi_1 \tag{21}$$

Because of the complexity of the objective function, it is unlikely that any closed form solution exists for the optimization problems. Therefore, we use a direct-search and derivative-free method called the Local Unimodal Sampling Algorithm to seek the optimal solution. A simple pseudo-code of the procedure (Perdersen 2010) is provided below, and the algorithm stops after 100 iterations. We also run the algorithm multiple times and compare the generated solution to significantly increase the chance of reaching a global (instead of local) optimum.

- Step 1: Initialize μ_1 with a random uniform position in $[0, R/\pi_1]$.
- Step 2: Set the initial sampling range $[0, R/\pi_1]$ to cover the entire search-space: $d = R/\pi_1$.

- Step 3: Until the termination criterion is met, repeat the following:
  - Step 3.1: Randomly generate $a \approx (-d, d)$.
  - Step 3.2: Let $\mu_1' = \mu_1 + a$; if μ_1' is within $[0, R/\pi_1]$, go to Step 3.3, otherwise go back to Step 3.1.
  - Step 3.3: If μ_1' improves the objective function, then move to the new position by setting $\mu_1 = \mu_1'$; otherwise, decrease the sampling-range by multiplying with a factor 0.5: $d = 0.5d$, and go to Step 3.1.
- Step 4: Now μ_1 holds the best-found solution, and μ_2 is obtained using Eq. (15).

3.2 Numerical examples and sensitivity analyses

Sensitivity analysis was performed to observe the effects of model parameters on optimal solutions. The interesting parameters include the arrival rate of type 1 patients λ_1 (the arrival rate of type 2 patients is fixed); the total available budget, R; the unit cost of service rates, π_1 and π_2; the transition rates between severity levels, q_{21} and q_{10}; and capacities C_1 and C_2. For simplicity, we assume that $C_1 = C_2$. The results for the optimization model **P1** are shown in Fig. 6.

First, we analyze the impact of the total available budget. Consider the following baseline parameter values: $R = 1, \lambda_2 = 1, \lambda_1 = 0.5, q_{21} = 0.2, q_{10} = 0.1, \pi_1 = 0.75, \pi_2 = 0.25$ and $C_1 = C_2 = 100$. Figure 6 plots the optimal service rates and the total death rate (our objective that we minimize) by changing one parameter at a time. We first conduct sensitivity analysis for **P1**.

Observation 1: When the objective is to minimize the total death rate, it is usually optimal to allocate more resources to the type 2 queue. There are a few of exceptions. One is when the budget is sufficiently large, more resources are allocated to the type 1 queue, such as in Fig. 6a when R is greater than 2. Note that when R is greater than 2 in the baseline parameter settings, service rates allocated to both queues could be greater than the arrival rates, respectively, which rarely happens in realistic disaster relief operations.

Our results are different from those of Gong and Batta (2006), which show that it is always optimal to allocate all the resources to the queue serving more severe patients, in order to minimize the weighted average number of patients in queues. In Gong and Batta (2006)'s model, no deterioration of patients' health condition is considered and it is implicitly assumed that service rates for different types of patients are the same. In contrast, we explicitly model the deterioration from mild injured patients to severely injured patients and to possible death, and assign different unit cost rates for service rates. Our results show that most, but not all, resources should be allocated to the queue of mild injured patients, which is also consistent with the humanitarian ethics that patients that are dying should not be completely ignored due to their low survival probability.

Observation 2: The optimal resource allocation plan is much more sensitive to the change of q_{21} than that of q_{10}. Our conjecture is that the increase in q_{10} may not be large enough to offset the cost differences between π_1 and π_2, and λ_1 is relatively small compared to λ_2, requiring fewer resources. Also shown in Fig. 6e, optimal resource allocation is more sensitive to capacities when both C_1 and C_2 are small. This is because the difference between two actual arrival rates is relatively small when the capacities are small due to large rejection of both queues.

Observation 3: Figure 6f shows that μ_1 increases in λ_1, approaches μ_2, but never exceed μ_2 (since it always costs more to treat type 1 patients).

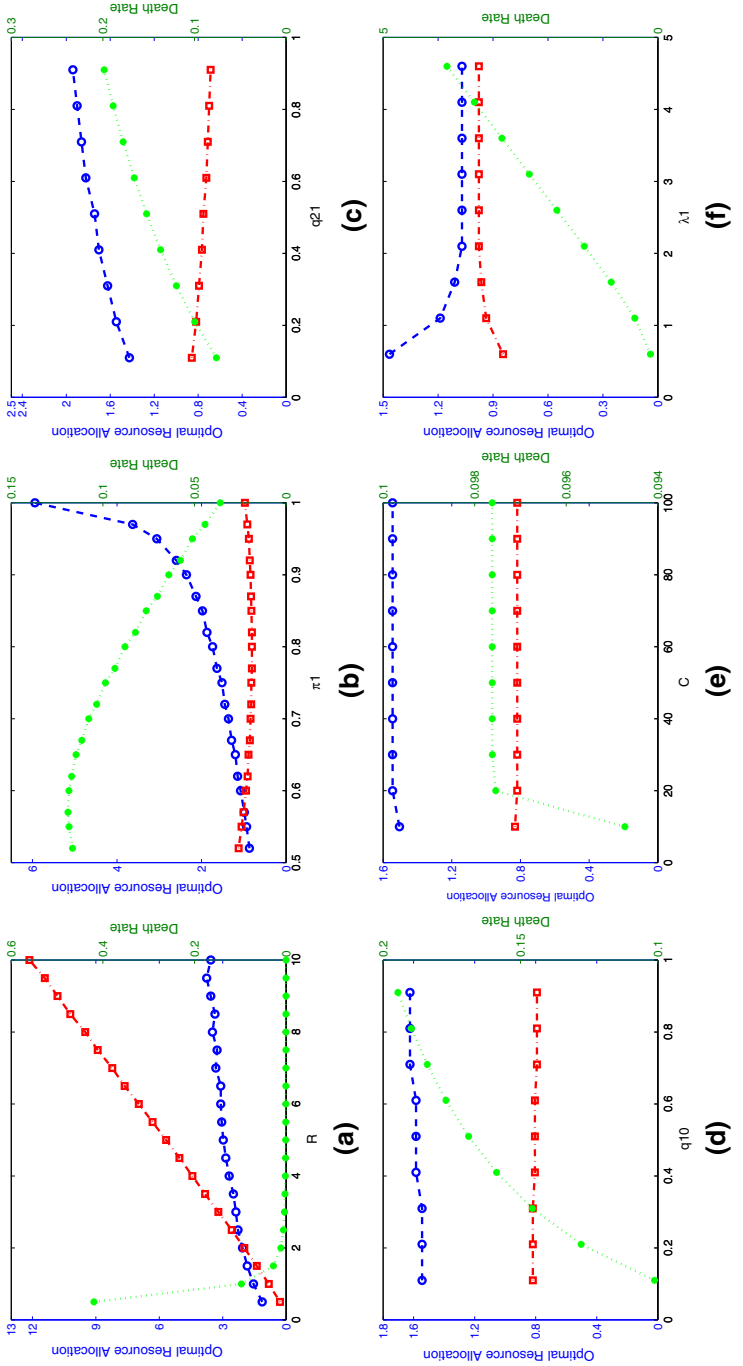

Fig. 6 Sensitivity analysis of the service rate and death rate (Model **P1**)

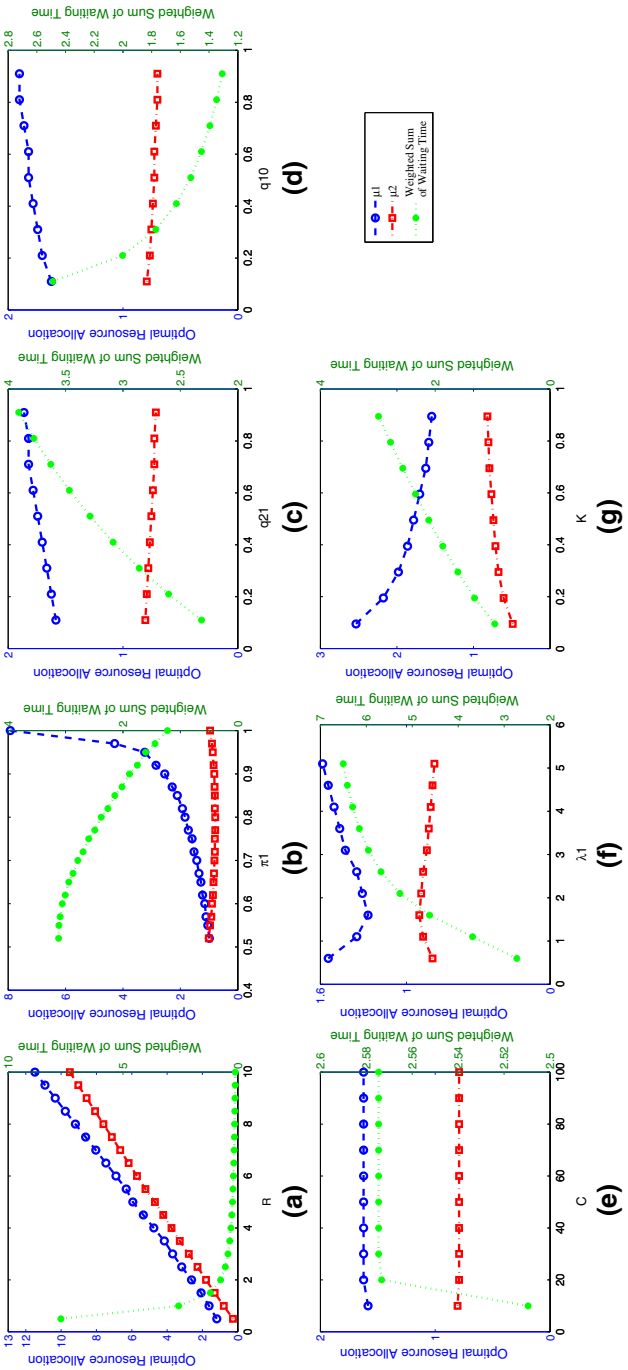

Fig. 7 Sensitivity analysis on weighted sum of expected system time (Model **P2**)

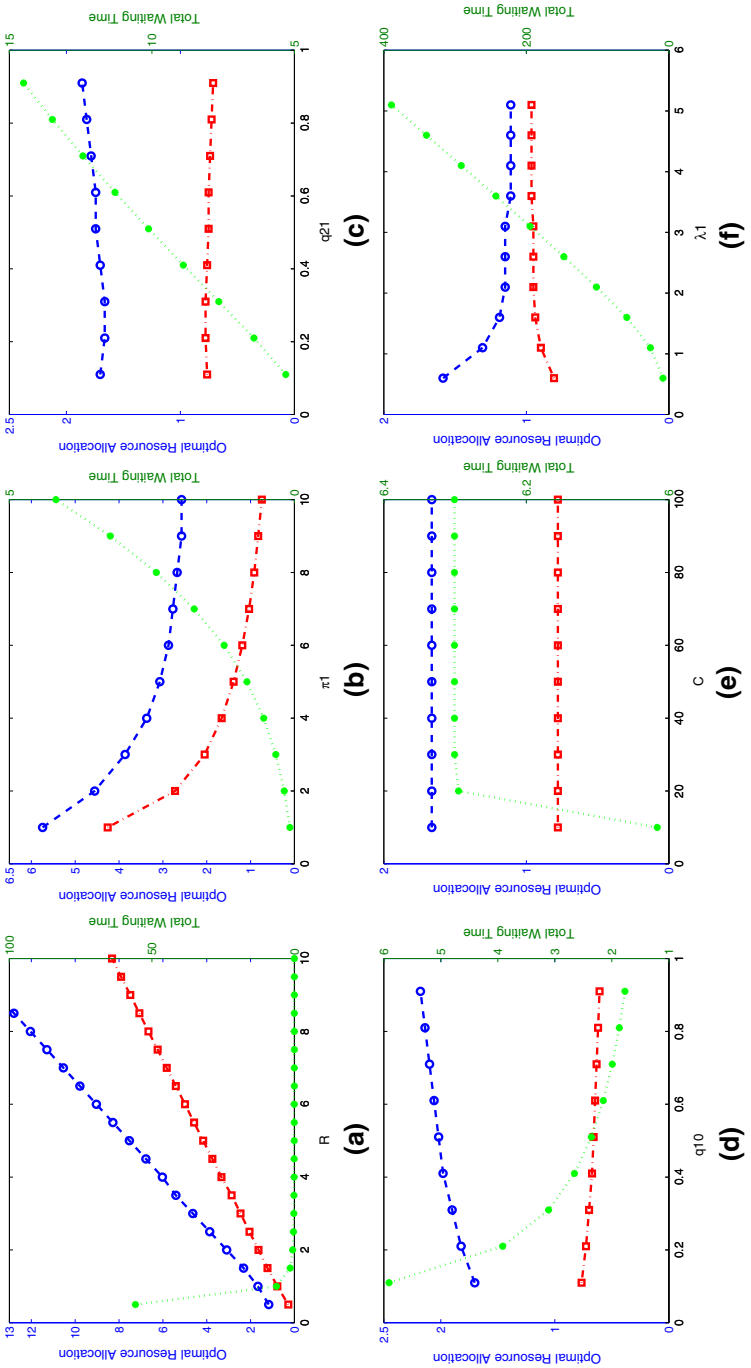

Fig. 8 Sensitivity analysis on the model of minimizing the total expected system time (Model **P3**)

We conduct similar numerical analysis for the models **P2** and **P3**, and analyze the effects of different parameter values on the system performance. Using the same baseline parameter settings of **P1**, the optimal medical resource allocation and the total weighted sum of the expected system time are shown in Figs. 7 and 8, for models **P2** and **P3**, respectively. One additional parameter K is needed for model **P2** in the baseline parameter settings and we let $K = 0.7$, since in practice, delaying patients in severe conditions could be worse than delaying the patients with mild conditions. We acknowledge that in other scenarios the time spent on type 2 patients could be more important, and the results are illustrated using the sensitivity analysis for K.

Observation 4: From Figs. 7 and 8, we observe that μ_2 is always larger than μ_1 under the objectives of minimizing the weighted average system time, and minimizing the total expected system time, respectively. Note that Figs. 7b, c, d, f and 8b, c, d, f show similar patterns as the ones in Fig. 6b, c, d, f. Figures 7e and 8e show that as the capacity increases, the optimal service rate allocated to the type 2 queue increases as opposed to the decreasing trend in Fig. 6e. This is because when the capacity increases, the actual arrival rates increase, and it is more efficient to give priority to type 2 patients. This also shows that different objectives may result in different qualitative insights on optimal service rates. However, one interesting observation is that regardless of the optimization criteria used (**P1**, **P2**, or **P3**), the optimal service rates converge to the same levels in Figs. 6e, 7e, and 8e.

Observation 5: When K increases, the optimal service rate allocated to the type 1 queue increases, but it still never exceeds μ_2. Even when $K = 1$, the optimal service rate allocated to the type 2 queue can not be zero due to deterioration. Note that all the observations are based on the assumption $C_1 = C_2$.

4 Conclusion and future research directions

In this paper, we present a novel model of allocating medical resources for disaster operations management, which provides novel "policy analytics" modeling tools and insights for governmental and non-governmental agencies in the health care field. In particular, we consider two-priority queues, each modeled as a single server to one type of dedicated victims. We allow for transitions from mild conditions to severe conditions. We provide both analytical solutions and numerical illustrations for this queueing network. Two optimization models are developed to minimize the total death rate and weighted sum of expected system time, respectively. A simple heuristic algorithm is used to find the optimal resource allocation plans. We also numerically illustrate the properties of the solutions.

There are several possible directions for future work. First, in our model, we assume that each patient type is served by a single server, and it would be interesting to consider multi-server scenarios. Secondly, it is assumed in this study that the arrival rates and service rates of the patients are known, and it would be valuable to derive them from data. Thirdly, we assume that each patient type has a designated queueing system, and the patients are rejected if that designated system is full. It would be interesting to study the scenario where both patient types share a common system, and how to allocate resources between patient types within such a common system. Fourthly, for analytical tractability, we only study the steady-state performance for the queueing system. This is valid for a scenarios where the arrival process is stable for a relatively long time period. It would be interesting to study the transient behavior for a scenarios when the arrival process is not homogeneous (e.g., a large amount of patients may arrive in a short period of time following disasters). Simulation and time-dependent fluid model could be used for studying such transient behavior. It would also be practical to

consider batch-arrivals. Additionally, it would be interesting to consider patient types that deteriorate in health but not necessarily end in death. Finally, since this paper focuses on model development rather than problem solving, we use a simple Local Unimodal Sampling Algorithm to seek optimal solutions, which are applicable for small-scale problems; it is worthwhile designing more efficient heuristic algorithms to solve large-scale problems.

Appendix 1

The solution method used for solving the 2-D Markov process in this paper involves decomposing the 2-D process into a set of 1-D Markov processes. Instead, we can obtain the numerical results by solving the steady-state equations for the 2-D Markov process directly. Table 1 provides the results comparing the decomposition versus the numerical methods in three examples where $C_1 = C_2 = 4, \lambda_2 = 1, q_{21} = 0.2, q_{10} = 0.1$: Example 1 ($\lambda_1 = 0.8, \mu_1 = 1.0, \mu_2 = 1.5$), Example 2 ($\lambda_1 = 0.5, \mu_1 = 0.5, \mu_2 = 1.5$), and Example 3 ($\lambda_1 = 0.8, \mu_1 = 1.0, \mu_2 = 1.5$). From Table 1 we observe that the comparison results are pretty good: the absolute errors are very small.

Further more, we study how such absolute errors change when the number of states C increases. For each of the three examples in Table 1, we extend to study $C = 5, 10, 15, 50$ as shown in Table 2. The results show that both the average and the standard deviation of the absolute errors (across states; it is not meaningful to report probabilities for each of the states $i = 1, C$ for each queue) decrease when C increases. This confirms that our approximation method is stable.

Table 1 Comparing analytical with approximating probabilities for each states and each queue

		Example 1			Example 2			Example 3		
	State	Analytical prob.	Approx prob.	Absolute error	Analytical prob.	Approx prob.	Absolute error	Analytical prob.	Approx prob.	Absolute error
Type 1 queue	0	0.2842	0.2928	0.0086	0.4945	0.5005	0.0060	0.4023	0.4517	0.0494
	1	0.2521	0.2508	0.0013	0.2735	0.2679	0.0056	0.2830	0.2484	0.0346
	2	0.2045	0.1996	0.0049	0.1388	0.1349	0.0039	0.1743	0.1442	0.0301
	3	0.1529	0.1496	0.0033	0.0649	0.0654	0.0005	0.0943	0.0924	0.0019
	4	0.1063	0.1072	0.0009	0.0282	0.0313	0.0031	0.0461	0.0632	0.0171
Type 2 queue	0	0.4190	0.4231	0.0041	0.5374	0.5381	0.0007	0.2602	0.2631	0.0029
	1	0.2793	0.2821	0.0028	0.2687	0.2690	0.0003	0.2602	0.2631	0.0029
	2	0.1666	0.1659	0.0007	0.1225	0.1223	0.0002	0.2187	0.2193	0.0006
	3	0.0902	0.0873	0.0029	0.0514	0.0510	0.0004	0.1591	0.1566	0.0025
	4	0.0448	0.0416	0.0032	0.0200	0.0196	0.0004	0.1019	0.0979	0.0040
Average				0.00327			0.00211			0.0146

Table 2 Average and standard deviations for absolute errors between analytical and approximating probabilities when C changes

	Example 1		Example 2		Example 3	
C	Average absolute error	Standard deviation for absolute error	Average absolute error	Standard aeviation for absolute error	Average absolute error	Standard deviation for absolute error
5	0.0033	0.0032	0.0069	0.0068	0.0038	0.0043
10	0.0031	0.0044	0.0065	0.0087	0.0030	0.0045
15	0.0024	0.0039	0.0049	0.0078	0.0022	0.0039
20	0.0019	0.0035	0.0037	0.0071	0.0017	0.0035
25	0.0015	0.0033	0.0030	0.0065	0.0013	0.0032
30	0.0013	0.0030	0.0025	0.0061	0.0011	0.0030
35	0.0011	0.0028	0.0022	0.0057	9.65E–04	0.0028
40	9.52E–04	0.0027	0.0019	0.0054	8.47E–04	0.0026
45	8.48E–04	0.0025	0.0017	0.0051	7.55E–04	0.0025
50	7.65E–04	0.0024	0.0015	0.0049	6.81E–04	0.0024

Appendix 2

To test the optimality of the Local Unimodal Sampling algorithm in Sect. 3.1, we randomly generate twelve instances of the medical resource allocation problem of our interest, and compare the results obtained from unimodal sampling algorithm against the direct grid-search method (exhaustive search). The grid-search method involves setting up a suitable grid in the design space, evaluating the objective function at all grid points, and finding the grid point corresponding to the lowest function value (Rao 2009). The reason for choosing the simple grid-search approach for comparison purpose is that it may not be safe to use approximation or heuristic methods that avoid doing an exhaustive parameter search. For the twelve problem instances we tested, the results (expected total death rate) from unimodal sampling algorithm are 0.18 % higher than those from the grid-search on average, and a more detailed comparison is shown in Fig. 9. The comparison results provide evidence that the Local Unimodal Sampling Algorithm is acceptable for the optimization problems studied in this paper.

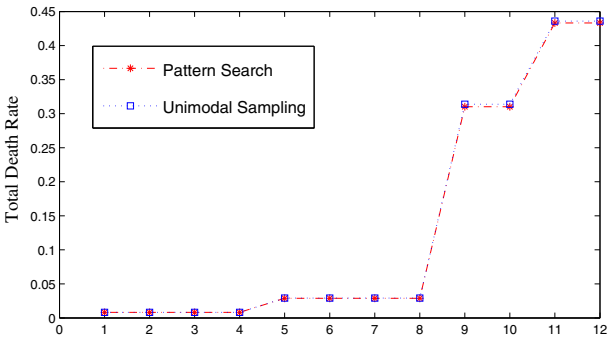

Fig. 9 Unimodal sampling algorithm versus direct grid-search

References

Altay, N., & Green, W. G. (2006). OR/MS research in disaster operations management. *European Journal of Operational Research*, *175*(1), 475–493.

Argon, N. T., Ding, L., Glazebrook, K. D., & Ziya, S. (2009). Dynamic routing of customers with general delay costs. *Probability in the Engineering and Informational Sciences*, *23*(2), 175–203.

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

Barbarosolu, G., Ozdamar, L., & Ceik, A. (2002). An interactive approach for hierarchical analysis of helicopter logistics in disaster relief operations. *European Journal of Operational Research*, *140*(1), 118–133.

Bostick, N. A., Subbarao, I., Burkle, F. M., Hsu, E. B., Armstrong, J. H., & James, J. J. (2008). Disaster triage systems for large-scale catastrophic events. *Disaster Medicine and Public Health Preparedness*, *2*(Suppl XX—-XX1), S35–39.

Cabrill, T. B. (1974). Optimal control of a maintenance system with variable service rates. *Operations Research*, *22*(4), 736–745.

Coles, J., Zhuang, J., & Yates, J. (2012). Case study in disaster relief: A descriptive analysis of agency partnerships in the aftermath of the January 12th, 2010 Haitian earthquake. *Socio-Economic Planning Sciences*, *46*(1), 67–77.

Cookson, R., & Dolan, P. (2000). Principles of justice in health care rationing. *Journal of Medical Ethics*, *26*(5), 323–329.

Dolan, P., & Cookson, R. (2000). A qualitative study of the extent to which health gain matters when choosing between groups of patients. *Health Policy*, *51*(1), 19–30.

Fiedrich, F., Gebauer, F., & Rickers, U. (2000). Optimized resource allocation for emergency response after earthquake disasters. *Safety Science*, *35*(1–3), 41–57.

George, J. M., & Harrison, M. J. (2001). Dynamic control of a queue with adjustable service rate. *Operations Research*, *49*(5), 720–731.

Glazebrook, K. D., Ansell, P. S., Dunn, R. T., & Lumley, R. R. (2004). On the optimal allocation of service to impatient tasks. *Journal of Applied Probability*, *41*(1), 51–72.

Gong, Q., & Batta, R. (2006). A queue-length cutoff model for a preemptive two-priority M/M/1 systems. *SIAM Journal on Applied Mathematics*, *67*(1), 99–115.

Gong, Q., & R. Batta. 2007. Allocation and reallocation of ambulances to casualty clusters in a disaster relief operation. IIE Transactions 39 27–39(13).

Haghani, A., & Oh, S. C. (1996). Formulation and solution of a multi-commodity, multi-modal network flow model for disaster relief operations. *Transportation Research Part A: Policy and Practice*, *30*(3), 231–250.

Hausken, K., & Zhuang, J. (2013). The impact of disaster on the interaction between company and government. *European Journal of Operational Research*, *225*(2), 363–376.

Hick, J. L., Barbera, J. A., & Kelen, G. D. (2009). Refining surge capacity: Conventional, contingency, and crisis capacity. *Disaster Medicine and Public Health Preparedness*, *3*(1), S59–67.

Insua, D., Ruggeri, F., & Wiper, M. (2012). Bayesian analysis of stochastic process models. New York: Wiley.

Jenkins, J. L., McCarthy, M. L., Sauer, L. M., Green, G. B., Stuart, S., Thomas, T. L., et al. (2008). Mass-casualty triage: Time for an evidence-based approach. *Prehospital and Disaster Medicine*, *23*(1), 3–8.

Kahn, C. A., Schultz, C. H., Miller, K. T., & Anderson, C. L. (2009). Does START triage work? An outcomes assessment after a disaster. *Annals of Emergency Medicine*, *54*(3), 424–430.

Li, D., & Glazebrook, K. D. (2010). An approximate dynamic programing approach to the development of heuristics for the scheduling of impatient jobs in a clearing system. *Naval Research Logistics*, *57*(3), 225–236.

Macario, A. (2010). Is it possible to predict how long a surgery will last? Medscape. Jul 14, 2010. http://www.medscape.com/viewarticle/724756. Accessed in November 2013

Ozdamar, L., Ekinci, E., & Kucukyazici, B. (2004). Emergency logistics planning in natural disasters. *Annals of Operations Research*, *129*(1–4), 217–245.

Perdersen, M. E. H. 2010. Tuning and simplifying heuristical optimization. Ph.D. thesis, University of Southampton, School of Engineering Sciences.

Rao, S. S. (2009). Engineering optimization: Theory and practice. Hoboken: Wiley.

Sacco, W. J., Navin, D. M., Fiedler, K. E., Waddell, R. K, I. I., Long, W. B., & Buckman, R. F. (2005). Precise formulation and evidence-based application of resource-constrained triage. *Academic Emergency Medicine*, *12*(8), 759–770.

Sheu, J. B. (2007). An emergency logistics distribution approach for quick response to urgent relief demand in disasters. *Transportation Research Part E: Logistics and Transportation Review*, *43*(6), 687–709.

Sheu, J. B. (2010). Dynamic relief-demand management for emergency logistics operations under large-scale disasters. *Transportation Research Part E: Logistics and Transportation Review*, *46*(1), 1–17.

Sinuff, T., Kahnamoui, K., Cook, D. J., Luce, J. M., & Levy, M. M. (2004). Rationing critical care beds: A systematic review. *Critical Care Medicine*, *32*(7), 1588–1597.

Super, G., Groth, S., & Hook, R. (1994). *START: Simple triage and rapid treatment plan*. Newport Beach, CA: Hoag Memorial Presbyterian Hospital.

Weber, R., & Stidham, S. (1987). Optimal control of service rates in networks of queues. *Advances in Applied Probability*, *19*(1), 202–218.

Xiong, W., & Altiok, T. (2009). An approximation for multi-server queues with deterministic reneging times. *Annals of Operations Research*, *172*(1), 143–151.

Yi, W., & Kumar, A. (2007). Ant colony optimization for disaster relief operations. *Transportation Research Part E: Logistics and Transportation Review*, *43*(6), 660–672.

Zhuang, J., & Bier, V. M. (2007). Balancing terrorism and natural disasters–Defensive strategy with endogenous attacker effort. *Operations Research*, *55*(5), 976–991.

Zhuang, J., G. Saxton, & H. Wu. (2014). Publicity vs. impact: A sequential game with a non-profit organization and N donors. *Annals of Operations Research, 221*(1), 469–491.