Optimizing Emergency Preparedness and Resource Utilization in Mass-Casualty Incidents

Panagiotis P. Repoussis

Department of Marketing & Communication, School of Business, Athens University of Economics & Business, Athens 11362, Greece; and School of Business, Stevens Institute of Technology, Hoboken, NJ 07030, USA

Dimitris C. Paraskevopoulos

School of Management, University of Bath, Claverton Down, Bath, BA2 7AY, UK

Alkiviadis Vazacopoulos

Optimization Direct Inc, Harrington Park, NJ 07640, USA; and School of Business, Stevens Institute of Technology, Hoboken, NJ 07030, USA

Nathaniel Hupert

Department of Public Health, Weil Medical College of Cornell University, New York, NY 10065, USA

Abstract

This paper presents a response model for the aftermath of a Mass-Casualty Incident (MCI) that can be used to provide operational guidance for regional emergency planning as well as to evaluate strategic preparedness plans. A mixed integer programming (MIP) formulation is proposed for the combined ambulance dispatching, patient-to-hospital assignment, and treatment ordering problem. The goal is to allocate effectively the limited resources during the response so as to improve patient outcomes, while the objectives are to minimize the overall response time and the total flow time required to treat all patients, in a hierarchical fashion. The model is solved via exact and MIP-

Email addresses: prepouss@stevens.edu (Panagiotis P. Repoussis), d.paraskevopoulos@bath.ac.uk (Dimitris C. Paraskevopoulos), alkis@optimizationdirect.com (Alkiviadis Vazacopoulos), nah2005@med.cornell.edu (Nathaniel Hupert)

based heuristic solution methods. The applicability of the model and the performance of the new methods are challenged on realistic MCI scenarios. We consider the hypothetical case of a terror attack at the New York Stock Exchange in Lower Manhattan with up to 150 trauma patients. We quantify the impact of capacity-based bottlenecks for both ambulances and available hospital beds. We also explore the trade-off between accessing remote hospitals for demand smoothing versus reduced ambulance transportation times.

Keywords: Emergency Medical Services; Mass-Casualty Incident; Triage; Scheduling; Ambulance Dispatching; Local Search; Resource Allocation

1. Introduction

Any medical incident in which casualties actually or potentially overwhelm local emergency response and hospital treatment capability may be termed a mass-casualty incident (MCI). These are typically major events, such as transportation accidents and terrorist bombings, with many casualties, though many jurisdictions define an MCI using a relatively small numerical threshold (e.g., 5 casualties from one incident in New York City while in South Korea 6 such casualties)(Arnold et al., 2004; Park et al., 2016). MCIs may overwhelm local treatment capability either due to sheer numbers of injured patients all needing treatment at the same time, or a potentially smaller number of patients who require advanced care (e.g., neurosurgical care) that is in relatively short supply locally. The International Institute for Counter-Terrorism has recorded over 33,000 terrorist incidents in the world since 1975, while lately the potential for terrorist activity is on the rise. Furthermore, the increasing frequency and severity of megastorms, such as Hurricane Sandy, has made natural-origin MCIs more likely. This paper is concerned with the development of a response model for the aftermath of an MCI that can be used to optimize resource utilization, to provide operational guidance for regional emergency planning, and to evaluate strategic preparedness plans.

As described by Mills et al. (2014) an MCI creates a sudden spike in demand for the emergency response resources within an area, and as a result, even patients who are in critical condition may not have timely access to these resources that are essential for their survival. During an MCI, it sometimes happens that only a limited number of ambulances are available to transport patients, requiring ambulances to make multiple trips from the MCI site(s) to the hospitals or forcing reliance on self-transportation (der

Heide, 2006). Dispatching software systems typically retrieve the locations and contact information for the hospitals nearest to the event, but this prioritizes the travel time over other factors required for optimal response, such as availability of existing medical resources at these facilities. Emergency medical service (EMS) systems are thus challenged during MCI response to allocate effectively a set of limited resources to the patients awaiting treatment and transportation to hospitals.

From the operational perspective, we need to determine which of the available hospitals should be included in the response according to their category, trauma level, capacity and proximity to the MCI site(s), how many ambulances should be utilized, where ambulances should transport each subsequent patient, and how many patients should be transported to each hospital. The arrival times of patients at the hospitals, the hospitals' throughput capability, and the patient treatment times will dictate the treatment order (scheduling) of the transported patients at each hospital and the time required to treat all patients. During the ambulance dispatching and patient-to-hospital assignment processes, we also need to follow a triage protocol as well as to match the specific treatment needs of the patients. On the other hand, given the potential location and size of an MCI we need to measure the simulated response efficiency and to identify the bottlenecks.

Ambulance dispatching decisions affect both the patient waiting times at the site and at the hospitals as well as the availability of ambulances. In practice, one commonly employed dispatch strategy is called "scoop-and-run", whereby patients are sent as quickly as possible to the closest hospital in order to minimize the dispatching times (this is standard practice in Israel, for example). However, this strategy ignores the specific needs of the patient, the triage protocol, and the current available capacity at the hospitals. For example, sending a large number of patients to the closest hospital may cause congestion resulting in long waiting times and unnecessary re-dispatching of patients in the worst case. As described by Carter et al. (1972) dispatching the closest idle ambulance to an emergency call is not always the optimal policy, if the objective is to minimize the response times.

Regardless the rich literature on emergency response problems, there remain no generally accepted, evidence-based guidelines to advise dispatchers on fundamental questions, such as which hospitals to include in a specific MCI response and how many casualties to transport to each. Ambulance dispatching has been performed mostly on the basis of the reliability and validity of EMS experts' cognitive abilities. It is possible, however, that dispatching

strategies using situational awareness information combined with knowledge of regional hospital capabilities, including destination facility-specific transportation times and treatment capability, could yield superior outcomes. For example, it is reasonable to expect that by balancing the load on the hospitals the level of care will be improved and the delays experienced by the patients will be reduced (Repoussis et al., 2015). Although divergent modeling approaches appear in the literature, it is likely that computerized models will be increasingly important in providing public accountability for the resource allocation decisions that have to be made in emergency situations.

The contribution of this paper is three-fold. First, we address the combined ambulance dispatching, patient-to-hospital assignment, and treatment ordering problem. In particular, we propose a rigorous mathematical formulation that captures all critical compatibility issues and prioritization aspects according to the Simple Triage and Rapid Treatment (START) triage protocol. Furthermore, we consider the makespan (*i.e.* the latest completion time) and the total flow time as hierarchical objectives. Second, we present a hybrid MIP-based construction heuristic and local search improvement algorithms that allow us to solve and find high quality solutions for otherwise computationally intractable large scale problem instances. Third, we study the effectiveness and efficiency of the new algorithms via a comprehensive study on randomly generated small- and medium-scale instances. We also try to identify how the availability of resources as well as the spatial and temporal characteristics affect the response times and the allocation of resources. Additionally, we demonstrate the applicability of the new model on an example MCI with realistic data. We consider the hypothetical case of a bombing at the New York Stock Exchange in Lower Manhattan. For a given number of ambulances we examine 3 scenarios, regarding the size of the MCI, with up to 150 patients. We lastly examine trade-offs between increasing the available capacity (e.g., adding hospital beds) of the hospitals nearby the site and including relatively distant hospitals instead.

The remainder of the paper is structured as follows. Section 2 briefly discusses the related work regarding models; 3 presents the mathematical model and discusses the combined ambulance dispatching, patient-to-hospital assignment, and treatment ordering problem; and Section 4 introduces the MIP-based construction heuristic and local search metaheuristic algorithms. Subsequently, Section 5 reports the computational experiments and results based on realistic data, and finally the paper concludes in Section 6.

2. Related Work

In practice, ambulance dispatching decisions are made in a dynamic environment; however, it is difficult to design and apply real-time dispatching tools because information is dynamic and often incomplete. Given this state of affairs, various strategic decisions can be evaluated a priori, such as, given what is known about regional surge capacity, which hospital should be included in regional disaster preparedness planning. Computer- or exercise-based modeling is therefore becoming increasingly important in providing a test-bed for the resource allocation decisions that have to be made in emergency situations, without the overt risk of harm to current patients.

Emergency vehicle deployment problems have been widely studied in the literature and various models and solution frameworks have been developed. A large part of this literature focuses on reducing dispatching response times in standard emergency call processes, with an emphasis on how to allocate emergency service stations and units. Toregas et al. (1971) proposed a location set covering model that minimized the number of ambulance required to cover all demand points with a preset coverage standard. The Hypercube Queueing model introduced by Larson (1974) was the first model to embed queueing theory in location problems. A survey on deterministic, stochastic and/or dynamic ambulance location and allocation models is provided by Brotcorne et al. (2003). More recent works in dynamic and real-time models for emergency vehicle dispatching and coverage relocation are those of Gendreau et al. (2006) and Haghani & Yang (2007). Interested readers may also refer to Boldberg (2004) for an overview on dispatching emergency service vehicles and to Bektas et al. (2014) for a detailed survey regarding models and algorithms for dynamic and stochastic vehicle dispatching problems.

Similarly, various models and decision support systems have been proposed for resource management in disaster response. The vast majority focuses on the location and allocation of emergency response units (Fiedrich et al., 2000) as well as on the supply and distribution of relief supplies (Barbarosoglu & Arda, 2004; Mete & Zabinsky, 2010). Few papers consider the transportation of casualties and flows of patients between locations (Wilson et al., 2013; Salman & Gul, 2014). Notably, many models assume that the same vehicles are used to distribute emergency supplies and simultaneously to transport casualties to treatment facilities (Yi & Kumar, 2007; Yi & Ozdamar, 2007; Ozdamar, 2011). In these works the main effort is to determine the flows of commodities and casualties between supply and de-

mand locations as well as the vehicle routes. On the other hand, Salman & Gul (2014) proposed a multi-period model to optimize capacity allocation and casualty transportation with the objective to minimize the traveling and waiting times as well as the cost of establishing new facilities. Lastly, an agent-based framework is presented by Bae et al. (2015).

Dynamic and robust multi-objective, multi-commodity, and multi-modal models for dispatching and routing vehicles in response to earthquakes are presented by Najafi et al. (2013, 2014). Their goal was to minimize hierarchically the transit and waiting times for transporting relief commodities and injured people. Barbarosoglu et al. (2002) presented a helicopter routing problem for collecting casualties. Chiu & Zheng (2007) addressed the evacuation problem in which multiple emergency responses and evacuation flow groups with different destinations and varying priorities coexist in the same traffic network. Lastly, Gong & Batta (2007) and Jotshi et al. (2009) considered the problem of dispatching ambulances to clusters of casualties. Data fusion was used to provide estimates for the problem entities and for clustering the casualty locations. The objective was to minimize the makespan, while the effect of allocation and re-allocation decisions was also considered.

Most ambulance dispatching papers described above address the problem of where and when ambulances should be located with the objective of minimizing response times. Although these response times are important, the ultimate clinical goal in MCI management is to have all patients stabilized and treated as early as possible in order to reduce morbidity and mortality. One main difference of our modeling framework compared to existing works is that we take into account both the dispatching response times and the treatment times–that is, hospital-based patient throughput times and their associated waiting times at the hospitals. Another difference is that we particularly keep track of each individual patient, instead of simplistically looking at the overall flows of patients between locations. Furthermore, in our formulation we assume that one patient is assigned to an ambulance at a time, rather than allowing bulk pickups at the same route. Since single-patient transport is the universal standard for such services, and bulk transport would require a change to what are now referred to as "crisis standards of care" (Hanfling et al., 2012), we felt it essential to base our model on actual daily practice patterns at least to establish a baseline for response parameterization.

Recently, Wilson et al. (2013) modeled the distribution of casualties to hospitals as a single-period multi-objective flexible job shop scheduling problem. Each patient is considered as a job and each responder unit as a ma-

chine. For each job there is a set of tasks, such as transport, pre-transport treatment, pre-rescue treatment, transportation to hospital and treatment. Each patient is assigned to a hospital and each responder is assigned a sequence of operations. Sequence dependent setup times occur when a responder unit moves between locations. The hospital's available capacity varies dynamically, and depends on either scheduled or self-transported patient arrivals. The authors adopted a multi-objective scheme for the evaluation of solutions, which considers the expected number of fatalities, the weighted total flow time, the appropriateness of hospital allocation, the responder idle times and the makespan. Overall, the problem is solved via a construction heuristic algorithm and a variable neighborhood descent metaheuristic algorithm. To our knowledge, this is the only work that provides a well-defined task scheduling framework at the level of individual patients; however, our approach provides a more comprehensive treatment of the entire response effort, including the scheduling of patients at the level of hospital beds.

Building on the above static model, Wilson et al. (2016) describe a real-time solution framework with continuous communication between the optimization model and problem environment. This allows key problem parameters (e.g. number of casualties, time required to complete key response tasks) to be updated dynamically. This new information is used to improve future predictions as well as to correct past errors. Simulation-based computational experiments show that the real-time framework improves the static approach (in terms of expected fatalities and suffering of casualties) and mitigates against poor communication speed. Earlier, a web-based simulation model for patient-to-hospital allocation from MCI locations has been proposed by Amram et al. (2012). The effort is to provide real-time information to the responders at the scene regarding driving times, trauma service level and the location of each hospital. Note that the above framework requires real-time capacity data, which is rarely available (i.e., few hospital bed management systems automatically feeding local or regional response databases).

Besides the actual allocation of resources, a number of papers, especially in the field of emergency medicine, are concerned with field triage (i.e., assessment of the health of each casualty and estimation of the extend of their injuries), patient-to-hospital allocation, and the on-site prioritization of patients for transportation to hospitals (Hupert et al., 2007). Recently, Mills et al. (2014) presented a fluid model of patient triage in MCI that considers resource limitations and the changes in survival probabilities over time. The proposed policies outperformed the START protocol in all simulated

scenarios. Note that the START protocol is widely adopted. First, given the variably injured (critical and non-critical) patients at each site, every patient is assigned a triage level. The patients are classified into four classes, i.e, minor, delayed, immediate and expectant. START gives the highest priority to patients in the intermediate class and the second highest to the delayed class. Once the MCI site is cleared from patients with time-dependent outcomes, patients from the minor and expectant classes are considered.

Dean & Nair (2014), Sacco et al. (2005), and Sung & Lee (2016) model the patient prioritization problem as an ambulance scheduling problem, while different rule-based triage schemes are evaluated to provide better response to the maximum number of patients. In particular, the objective function used by Sung & Lee (2016) aims to maximize the total expected number of survivors. The decision variable is the order of transportation and destination hospital for each patient. A set-partitioning formulation is proposed and solved via a column generation approach. Similar to the work of Mills et al. (2014) one finding is that the delayed-first rule outperforms the immediate-first rule in most cases other than low-workload, optimistic scenario cases.

This may be why authors like Zoraster et al. (2007), who reviewed the START protocol followed after a train crash, have concluded that large EMS systems must plan a priori the distribution of critically injured patients accruing at various locations to optimize the care provided. A corollary of this is that if trauma resources are at risk of being overwhelmed, there should be pre-established plans to address the risk of patient maldistribution. To date most MCI research focuses on the management of surge capacity and the ability to quickly add resources in the time of emergency (Amram et al., 2012). Notably, little attention is given on how to prevent or delay surges by better directing the flows of patients (e.g. avoid overtriage) and how to determine the minimum treatment capability needed to successfully treat the casualties of an MCI. Our paper seeks to incorporate these complicated issues and gaps into the foundations of a decision-support tool, and therefore presents an analytical approach and quantitative framework that can be used to generate and evaluate realistic emergency preparedness plans.

3. Mathematical Model

3.1. Problem Description & Operational Realties

Similar to the events sequence discussed by Fitzsimmons (1973) for standard EMS calls, Figure 1 depicts the sequence of events associated with an

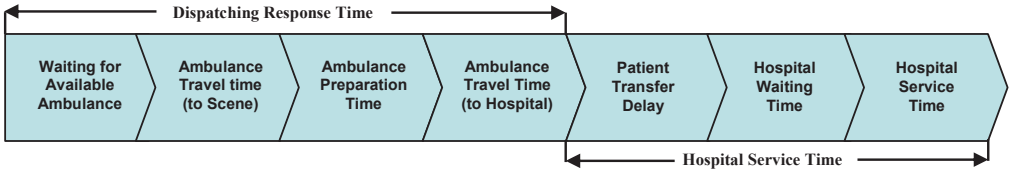


Figure 1: Sequence of events during an MCI

MCI. The first phase is related to the dispatching of ambulances and the transportation of patients to hospital facilities. Besides the travel time components, it is also important to consider delays related to ambulance preparation time for the subsequent patient (either at the hospital after drop-off or in the field). The second phase is related to the processing and treatment times within the hospitals.

At the MCI site(s) patients await initial assessment (which may involve acute medical treatment), triage, and transportation to hospital facilities. Although it depends on the scale of the event, it is reasonable to consider that the number of ambulances (or responders) available is limited compared to the demand. Therefore, the ambulances are required to make multiple trips from the sites to the hospitals. To the other end, the hospitals have different characteristics with regard to their trauma "level" (best thought of as capabilities), their emergency department capacity (currently referred to as part of their "immediate bed availability"), the treatment times that can be anticipated for traumatically injured or otherwise affected patient (which can be represented as patient throughput) and the distances from the MCI site(s) to participating hospitals. In our formulation, the treatment capability of each hospital is measured based on the number of staffed and available emergency department (ED) trauma bays with associated operating rooms (OR) that are recycled according to critical or non-critical processing times. The treatment times are different for every patient and for every hospital, and refer to standard evidence-based patient management times.

Regarding decision making processes and specifically triage, a fixed-priority ordering scheme is adopted among different classes of patients, based on the standard START protocol mentioned earlier. Regarding treatment order, the patients are processed sequentially, and all hospital resources (EDs and ORs) are recycled after each use. Without loss of generality, it is reasonable to consider that the treatment sequence is dictated by the within-class

patient arrival times in a first-come first-served (FIFO) basis. Note that in the special case of a single MCI site, the FIFO treatment order protocol is consistent with the initial triage assigned at the site.

Other elements that restrict the patient-to-hospital assignments are the hospital's capabilities, its trauma level, and its total bed capacity. Particularly, a patient may suffer from a injury type that requires treatment at a hospital capable of providing the specialized care (e.g., burn care). Another restriction is that high priority immediate patients can only be transported to trauma level I and II hospitals. Lastly, the number of beds sets the upper limit on the number of the patients that can be assigned to a hospital.

All data regarding the hospital's overall capacity and expected treatment capability are known in advance; this data can be collected from a variety of sources. Although different types of ambulances may be involved (e.g. Basic and Advance Life Support units), we consider only one type and we assume that each ambulance carries one patient. The number of available ambulances throughout the response effort is known; however, a maximum limit on the number of trips is imposed for each ambulance. The traveling times are known, and they are proportional to the geographical distances between hospitals and sites. The latter is a reasonable assumption that it is often made in practice; however, it is also often observed that during disasters ambulance travel times in urban environments with severe congestion may have a non-linear relationship with the distances. Readers may refer to Budge et al. (2010) for an empirical analysis on ambulance travel times.

In keeping with standard views about the urgency of emergency care (i.e., the commonsensical "golden hour" concept), long delays with respect to the completion of treatment for a given patient leads to a higher mortality rate in our framework. For this reason, we consider the minimization of the latest completion time (i.e., makespan) and the total flow time for all patients as hierarchical objectives. The completion time includes the waiting times for transportation at the site, the dispatching response times (traveling time plus loading and unloading of the patient at the ambulance), waiting times for treatment at the hospital, and the treatment times at the hospital.

3.2. Model Formulation

Let an undirected graph $G = (V, A)$ represent the transportation network, which spans all geographical areas of interest. The set of nodes V refers to the disaster sites, and the hospital facilities. In particular, let $H = \{1, \ldots, n_h\}$ denote the subset of hospitals, and $S = \{1, \ldots, n_s\}$ the subset of MCI sites

(such that $V = S \cup H$), where n_h is the total number of hospitals and n_s is the total number of sites. Let $A = \{(s, h) \in S \times H : s \neq h\}$ refer to the set of arcs. Whenever an ambulance traverses an arc $(s, h) \in A$ a travel time $d_{sh} \in \mathcal{R}_+$ incurs. Without loss of generality, it is assumed that the travel time matrix $[d_{sh}]$ is symmetric and satisfies the triangle inequality.

Considering that the patients can be seen as a set of jobs (see Wilson et al. (2013) for a similar approach), the proposed model formulation follows the representation of a Flexible Job Shop Scheduling Problem (FJSP)(Rocha et al., 2008) with unrelated parallel machines and sequence- and machine-dependent setup times. The FJSP consists of programming several jobs to be processed by several parallel identical and/or unrelated machines. Each job should be scheduled to a specific machine and the order in which each machine will process its jobs should be decided. The processing times of each job depends on the machine and there is also a sequence dependent setup time whenever a machine finishes processing a job.

Let $P = \{1, \ldots, n_p\}$ denote the set of patients (jobs) and $R = \{1, \ldots, n_r\}$ denote the set of ambulances, where n_p is the total number of patients and n_r is the total number of ambulances. Each job P_u, $1 \leq u \leq n_p$, consists of a sequence of 2 ordered operations/tasks, i.e., $o_{u,1}$ (transportation to a hospital) and $o_{u,2}$ (treatment at the hospital). Assuming that the ambulances and hospitals represent the set of machines, each operation $o_{u,1}$ can be processed by any ambulance $r \in R$, $1 \leq r \leq n_r$, and each operation $o_{u,2}$ can be processed by any hospital $h \in H$, $1 \leq h \leq n_h$, unless otherwise stated.

Regarding the first set of operations, the processing time $l_{u,1}$ is known for every patient u and all ambulances are identical. The processing time equals to the loading plus the unloading times of a patient at the ambulance. Also, the traveling times are incorporated as sequence-dependent setup times, since they vary according from which site the patient in question needs to be picked up, and to which hospital has been assigned. A binary indicator I_u^s is used to indicate the site s where the patient u is located. In contrast, the patient treatment times depend on the hospital (which here is an unrelated parallel machine). In particular, each patient has different treatment (processing) times $l_{u,2}^h$ at each hospital h. A binary indicator I_u^h is used to indicate whether a hospital $h \in H$ can provide the type of care the patient u needs.

Let T_s denote the disaster time at each site $s \in S$. All ambulances are assumed to be available at the time of disaster; however, we consider that there is a response delay T_D to account for the time needed by the first responders to arrive and to perform the field triage. An ambulance can

process at most one operation at a time and preemption is not allowed. Let $J_r = \{1, \ldots, n_j^r\}$ denote the jobs/trips order. Each ambulance r can perform a limited number of return trips n_j^r. The processing characteristics of the hospitals have a similar setup, in that we let $B_h = \{1, \ldots, n_b^h\}$ denote the treatment order of hospital h, where n_b^h is the total number of beds (capacity).

Triage information k_p is assumed to be available for each patient p to indicate the triage level. Overall, two priority levels are considered as described above. The higher the triage level of a patient is, the greater is his/her priority compared to other patients at the same MCI site.

Based on the above representation, the solution of the examined problem must include ordered lists of tasks to be allocated to each ambulance and hospital. Following the formulations that appear in the earlier work of Rocha et al. (2008) for the FJSP, we define two sets of binary variables with discrete positions in the processing sequence for each ordered list of tasks as follows

$$x_{rp}^{(j)} = \begin{cases} 1 & \text{if patient } p \text{ assigned to ambulance } r \text{ on its } j\text{th trip} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

$$z_{ph}^{(b)} = \begin{cases} 1 & \text{if patient } p \text{ is assigned at the } b\text{th bed at hospital } h \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Due to the the sequence-dependent setup times for each ambulance (i.e., return trips from the hospitals to the site), we define another set of "flow like" binary variables to track the sequence of hospitals visited by each ambulance:

$$y_{rh}^{(j)} = \begin{cases} 1 & \text{if ambulance } r \text{ on its } j\text{th trip arrive at hospital } h \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

In addition to the above binary variables, a set of continuous and non-negative decision and auxiliary variables are defined to capture the time stamps of the model during the transportation and treatment of patients:

- a_{rj}: Arrival time of ambulance r at the site for performing the jth job
- at'_{rj}: Transportation time of ambulance r for job j (load, unload and travel time to a hospital)
- at''_{rj}: Deadhead travel time of ambulance r for job j (return trip to the site from a hospital)

- st_{hb}: Treatment start time of the patient assigned to bed b at hospital h

- w_p: Waiting time for transportation at the site for patient p

- aw^b_{ph}: Mapping of waiting time w_p of patient $p \in P$ assigned at hospital $h \in H$ at bed $b \in B_h$

The objective is to minimize the makespan, i.e., the maximum of the completion time of treatment of all patients, and it is denoted as C_{max}.

$$\underset{x,y,z,a,at,w,aw,st}{\text{minimize}} \; C_{max} \tag{4}$$

This is subject to the following sets of constraints. The first set of constraints is related to the ambulance dispatching and the transportation of patients from the sites to the hospitals.

$$\sum_{r \in R} \sum_{j \in J_r} x^j_{rp} = 1 \quad \forall p \in P \tag{5}$$

$$\sum_{p \in P} x^j_{rp} \leq 1 \quad \forall r \in R, j \in J_r \tag{6}$$

$$\sum_{h \in H} y^j_{rh} \leq 1 \quad \forall r \in R, j \in J_r \tag{7}$$

$$\sum_{h \in H} y^j_{rh} = \sum_{p \in P} x^j_{rp} \quad \forall r \in R, j \in J_r \tag{8}$$

$$\sum_{p \in P} x^{(j-1)}_{rp} \geq \sum_{p \in P} x^j_{rp} \quad \forall r \in R, j \in J_r \setminus \{1\} \tag{9}$$

Constraint (5) ensures that each patient has been assigned with exactly one ambulance. Constraints (6) and (7) dictate that the ambulance carries at most one patient and visits at most one hospital at each trip. This set of constraints can be easily generalized to capture cases where an ambulance can carry more than one and at most a maximum number of patients for each job, limited by its capacity. Constraint (8) links the assignment with the patient flow binary variables. Constraint (9) dictates that there must be a patient at position $j-1$ if there is another patient assigned at position j on the same ambulance for $j \geq 2$.

The second set of constraints control the time stamps.

$$a_{r(1)} \geq \left(\sum_{s \in S} I_p^s T_s + T_D \right) x_{rp}^{(1)} \quad \forall r \in R, p \in P \tag{10}$$

$$a_{rj} \geq a_{r(j-1)} + at'_{r(j-1)} + at''_{r(j-1)} \quad \forall r \in R, j \in J_r \setminus \{1\} \tag{11}$$

$$at'_{rj} \geq l_{p,1} x_{rp}^j + \sum_{h \in H} \left(\sum_{s \in S} I_p^s d_{sh} \right) y_{rh}^j - M(1 - x_{rp}^j) \quad \forall r \in R, j \in J_r, p \in P \tag{12}$$

$$at''_{rj} \geq \sum_{h \in H} \left(\sum_{s \in S} I_p^s d_{sh} \right) y_{rh}^j - G\left(1 - x_{rp}^{(j+1)}\right) \quad \forall r \in R, j \in J_r \setminus \left\{n_j^r\right\}, p \in P \tag{13}$$

Constraint (10) initializes the arrival time for every patient at the first position on each ambulance and sets the time greater or equal to the disaster time plus the initial response and field triage delay. Constraint 11 determines the arrival times of ambulances at the site(s) for the jobs. These arrival times can be seen as the processing start times, which are equal to the processing start times of the jobs at the previous positions adding the actual processing time and the setup time between two positions. The former is captured by Constraint (12) and refers to the loading, unloading and travel time from site to hospital. The latter is expressed by Constraint (13) and refers to the travel time from the hospital of the previous job to the site of the next assigned patient. The big M and G scalars can be calculated as follows:

$$M = \max_{s,h \in A} \{d_{sh}\} + \max_{p \in P} \{l_{p,1}\} \tag{14}$$

$$G = \max_{s,h \in A} \{d_{sh}\} \tag{15}$$

The third set of constraints is related to the treatment of patients.

$$\sum_{h \in H} \sum_{b \in B_h} z_{ph}^b = 1 \quad \forall p \in P \tag{16}$$

$$\sum_{p \in P} z_{ph}^b \leq 1 \quad \forall h \in H, b \in B_h \tag{17}$$

$$\sum_{p\in P} z_{ph}^{(b-1)} \geq \sum_{p\in P} z_{ph}^{b} \quad \forall h \in H, b \in B_h \setminus \{1\} \tag{18}$$

$$\sum_{b\in B_h} z_{ph}^{b} \leq I_p^h \quad \forall h \in H, p \in P \tag{19}$$

Constraint (16) denotes that each patient is assigned to exactly one hospital. Constraint (17) dictates that at most one patient can occupy a hospital bed. Constraint (18) ensures that if a patient is assigned at position b (bed) there is another patient waiting treatment at the previous position b-1 at the same hospital (starting from $2nd$ position). Constraint (19) maintains the compatibility of the patient-to-hospital assignments according to the patient priorities, injury types and hospital's trauma level.

The fourth set links hospital treatment and transportation times.

$$st_{h(b+1)} \geq st_{hb} + \sum_{p\in P} l_{p,2}^h z_{ph}^b \quad \forall h \in H, b \in B_h \setminus \left\{n_b^h\right\} \tag{20}$$

$$w_p \geq a_{rj} - \sum_{s\in S} I_p^s T_s - Q\left(1 - x_{rp}^j\right) \quad \forall r \in R, j \in J_r, p \in P \tag{21}$$

$$aw_{ph}^b \geq w_p - K\left(1 - z_{ph}^b\right) \quad \forall p \in P, h \in H, b \in B_h \tag{22}$$

$$st_{hb} \geq \sum_{p\in P} aw_{ph}^b + \sum_{p\in P}\left(\sum_{s\in S} I_p^s T_s + l_{p,1} + \sum_{s\in S} I_p^s d_{sh}\right) z_{ph}^b \quad \forall h \in H, b \in B_h \tag{23}$$

$$st_{h(|B_h|)} + \sum_{p\in P} l_{p,2}^h z_{ph}^{(|B_h|)} \leq C_{max} \quad \forall h \in H \tag{24}$$

$$w_{p'} \geq w_p \quad \forall p, p' \in P : p \neq p', \exists s \in S | I_p^s = I_{p'}^s = 1, k_p > k_{p'} \tag{25}$$

$$st_{hb} \geq 0, w_p \geq T_D, aw_{ph}^b \geq T_D \quad \forall p \in P, h \in H, b \in B_h \tag{26}$$

$$a_{rj} \geq 0, at'_{rj} \geq 0, at''_{rj} \geq 0 \quad \forall r \in R, j \in J_r \tag{27}$$

$$x_{rp}^j, y_{rh}^j, z_{ph}^b \in \{0,1\} \quad \forall r \in R, j \in J_r, p \in P, h \in H, b \in B_h \tag{28}$$

Constraints (20) updates the treatment start times at each position according to the treatment time of the patient assigned to the bed of at the previous position. Constraint (21) is used to determine the waiting time for transportation at the site for each patient. Note that the patient's waiting time at the site should always be greater than or equal to the initial response

delay T_D (see Constraint 26). Similarly, Constraint (22) carries on the information regarding the patient's waiting times at the level of hospital beds via the set of auxiliary variables aw. Q and K scalars can be calculated as:

$$Q = (|P| - 1)(M + G) \tag{29}$$

$$K = Q + \sum_{s \in S} I_p^s T_s + T_D \tag{30}$$

Constraint (23) is used to calculate the arrival times of patients at the hospitals and to determine the earliest start times for treatment. Constraint (24) dictates that the completion time of treatment at the last position $|B_h|$ of any hospital $h \in H$ should be less than or equal to the C_{max}. Note that the start times of empty hospital beds (if any) are equal to those of the last treated patient. Constraint (25) ensures that the triage protocol is respected among the patients at each site. Lastly, Constraints (26) to (28) impose non-negative bounds and binary restrictions, respectively.

The formulation (4) to (30) best fits our combined ambulance dispatching and patient treatment ordering problem. This is an $\mathcal{NP}$-hard combinatorial optimization model that requires substantial computational effort for determining optimal and/or near optimal feasible solutions even for small problems. Although this is a strategic and not an operational problem, heuristic and metaheuristic algorithms can be applied to obtain high quality solutions in reasonable computational times, assuming practical sizes, while implicit enumeration schemes can be used mainly to produce lower bounds. It is practically important that even non-operational (i.e., planning) emergency response models work towards the ability to produce relatively rapid results, since they are increasingly used in real-world planning settings such as tabletop exercises that require relative swift feedback cycles.

Finally, it is worth highlighting that the proposed model discretizes the positions in the processing sequences. This adds more information to the x, y and z variables. This is not necessary for capturing the operational realties of the problem. However, as described by Rocha et al. (2008), although the number of variables increases, this type of discretization is expected to produce tighter lower bounds during resolution, and to be faster on large instances, compared to other mathematical formulations.

3.3. Symmetry Breaking Constraints

It should be noted that Formulation (4) to (30) exhibits some degree of isomorphism. In particular, although the x-variables correctly indicate

the consistent assignment of ambulances to patients at optimality, they do so in an arbitrary fashion. Let for notational convenience x_{pr} denote the assignment of a patient p to an ambulance r without taking into account the processing position. For any optimal solution $X_\alpha = [x_{p1}\ x_{p2}\ \cdots\ x_{pr}]$, there exists another equivalent solution $X_\beta = \left[x_{pr}\ x_{p(r-1)}\ \cdots\ x_{p1}\right]$; that is, a solution where the labels of the –otherwise similar– ambulances have been reversed. In fact there exist a total of $r!$ such equivalent solutions accounting for all possible permutations of the r columns (ambulances).

Such isomorphism is undesirable, as it retards the rate of node-pruning during branch-and-bound. It can be eliminated with symmetry-breaking constraints that impose a lexicographic ordering among the ambulances. We declare one of the possible permutations as *nominal* and we prefer this permutation over all its isomorphic ones. In particular, we prefer the permutation that adheres to the following simple rule: *for all $r \in R$, ambulance r is the one that serves the patient with the smallest index out of those that are not served by any ambulance r', such that $r' < r$.* For example, under the effect of this rule, patient $p = 1$ is always served by ambulance $r = 1$; patient $p = 2$ is served either by ambulance $r = 1$, if it is to be served by the same ambulance as patient $p = 1$, or by ambulance $r = 2$, if it is to be served by a separate ambulance; for patient $p = 3$, three cases are possible: (i) it is served by ambulance $r = 1$, along with patient $p = 1$, (ii) it is served by ambulance $r = 2$, along with patient $p = 2$ and separately from patient $p = 1$, or separately from both patients $p = 1$ and $p = 2$ that are served together by ambulance $r = 1$, (iii) it is served by ambulance $r = 3$, otherwise.

The following set of constraints allows only solutions that correspond to nominal ambulance permutations. Note also that they dominate –and thus should replace– the set of Constraints (5) in Formulation (4) to (30).

$$\begin{aligned} &\sum_{r=1}^{p}\sum_{j=1}^{n_j^r} x_{rp}^j = 1 && \forall\, p \in P \\ &\sum_{j=1}^{n_j^r} x_{rp}^j \leq \sum_{p'=r-1}^{p-1}\sum_{j=1}^{n_j^r} x_{(k-1)p'}^j && \forall\, r \in R,\ r \geq 3,\ \forall\, p \in P,\ p \geq r. \end{aligned} \tag{31}$$

3.4. Valid Inequalities, Lower Bounds and Special Cases

Additional constraints can be also introduced in the model. Their effect with respect to the LP relaxation is not always strong. Furthermore, their addition may further delay the overall execution, given that the overall basis

is also increased. However, they may also improve convergence in some cases by helping the relaxed problem to find solutions that are close to optimal.

Specifically, in addition to Constraints (19) that ensure the compatibility of patient-to-hospital assignments, the following set of inequalities are also valid, and further strengthen the link between x and y binary variables:

$$x_{rp}^{j} + y_{rh}^{j} \leq 1 + I_{p}^{h} \quad \forall t \in T, j \in J_t, h \in H, p \in P \tag{32}$$

Another inequality that links x, y and z binary variables all together in a rather unique way is the following:

$$y_{rh}^{j} + 1 \leq \sum_{b \in B_h} z_{ph}^{b} + x_{rp}^{j} \quad \forall t \in T, j \in J_t \tag{33}$$

Similar to Constraints (9) and (18), we contend that it is a valid restriction that a hospital is visited at position $j-1$ only if there is a hospital visited at position j of the task list on the same ambulance for $j \geq 2$.

$$\sum_{h \in H} y_{rh}^{(j-1)} \geq \sum_{h \in H} y_{rh}^{j} \quad \forall r \in R, j \in J_r \setminus \{1\} \tag{34}$$

Lastly, the following set of inequalities can be added into the model to reduce its inherent degeneracy.

$$at'_{rj} \leq l_{p,1} x_{rp}^{j} + \sum_{h \in H} \left(\sum_{s \in S} I_{p}^{s} d_{sh} \right) y_{rh}^{j} + M(1 - x_{rp}^{j}) \quad \forall r \in R, j \in J_r, p \in P \tag{35}$$

$$at''_{rj} \leq \sum_{h \in H} \left(\sum_{s \in S} I_{p}^{s} d_{sh} \right) y_{rh}^{j} + G\left(1 - x_{rp}^{(j+1)}\right) \quad \forall r \in R, j \in J_r \setminus \left\{n_j^r\right\}, p \in P \tag{36}$$

$$w_p \leq a_{rj} - \sum_{s \in S} I_{p}^{s} T_s + Q\left(1 - x_{rp}^{j}\right) \quad \forall r \in R, j \in J_r, p \in P \tag{37}$$

$$aw_{ph}^{b} \leq w_p + K\left(1 - z_{ph}^{b}\right) \quad \forall p \in P, h \in H, b \in B_h \tag{38}$$

A trivial lower bound for the makespan can be computed by adding the initial waiting time for transportation, the minimum transportation and ambulance processing time, and the minimum treatment time at the hospital.

$$LB^{trivial} = T_D + \min_{s \in S} \left\{ I_{p}^{s} T_s \right\} + \min_{s,h \in A; p \in P} \left\{ (d_{sh} + l_{p,1}) \right\} + \min_{p \in P; h \in H} \left\{ I_{p}^{h} l_{p,2}^{h} \right\} \tag{39}$$

More tight bounds on the makespan can be found by adding the initial waiting time for transportation at the site, the minimum processing, transportation and waiting times of all patients times the smallest number of setups necessary for each machine, and the smallest total treatment time for all patients at the available hospitals.

$$LB^t = LB^{t1} + \max\left(LB^{t2}, LB^{t3}\right) \tag{40}$$

$$LB^{t1} = T_D + \min_{s \in S}\left\{I_p^s T_s\right\} + \frac{n_p}{n_r}\left(\min_{s,h \in A; p \in P}\left\{(d_{sh} + l_{p,1})\right\}\right) \tag{41}$$

$$LB^{t2} = \frac{n_r}{n_h}\left(\min_{p \in P; h \in H}\left\{I_p^h l_{p,2}^h\right\}\right) \tag{42}$$

$$LB^{t3} = \frac{1}{n_h}\sum_{p \in P}\min_{h \in H} l_{p,2}^h \tag{43}$$

As described earlier, a special case of the proposed model is the single MCI site. In this case, the mathematical model can be simplified. In particular, the y variables can be dropped, together with constraints (7) and (8), while constraints (12) and (13) can be re-written considering only z and x variables.

3.5. Hierarchical Objectives

So far, we have assumed the minimization of makespan as the objective of the problem (representing the overall response time). However, a key non-modifiable factor that this objective does not take into account is the time-dependent mortality of critically injured patients. The survival probability of critical injured patients it is typically modeled as an exponential (or linear in the best case scenario) function of the waiting time until treatment (Hupert et al., 2007). Therefore, a more rigorous way to measure the quality and efficiency of the overall response is to consider not only the makespan, but also the impact that this has on the mortality rate of critical patients.

One way to capture the mortality rate is to consider the minimization of the (weighted) total flow time F_w for all patients (Wilson et al., 2013). The later can be calculated as the sum of the (weighted) completion times of treatment of each patient. In the case different weights w_p are assumed for each patient p, then these weights should reflect the prioritization as defined by the triage level of each patient.

$$F_w = \sum_{p \in P} \sum_{h \in H} \sum_{b \in B_h} w_p z_{ph}^b ct_{hb} \tag{44}$$

The bilinear terms in (44) can be easily linearized using standard convexification techniques, such as big-M, convex hull, and indicator reformulations.

Another alternative is to adopt a hierarchical (a.k.a. lexicographical) bi-criteria optimization scheme (Hoogeveen, 2005). In our case either the makespan or the total flow time is considered as the (dominant) primary minimization objective f, and the other one is considered as the secondary objective g, respectively. This implies that at the first stage priority is given to minimize f and to find the optimal value f^*, whereas at the second stage, g is minimized subject to the additional constraint that $f \leq f^*$.

Finding the optimal solution for the hierarchical bi-criteria optimization problem via exact MIP approaches it is straightforward; however, it requires a multi-step solution approach. In particular, one needs to solve sequentially the corresponding single objective optimization problems, i.e., to find the optimum for the first stage, and subsequently to find among the set of optimal schedules for the primary objective the one that performs best for the secondary objective. In such hierarchical multi-objective settings, metaheuristic solution approaches can be more advantageous and more flexible, since they can regulate heuristically the search directions in multiple dimensions. To that end, it is worth highlighting that in case no criterion is dominant, then the above hierarchical optimization scheme may lead to a schedule that is unbalanced, i.e., the score on the secondary criterion can be greatly improved by compromising only a little on the first criterion.

In the remainder of the paper, we will use the notation $f|g$ to indicate the hierarchy of objectives. For example, the $C_{max}|F_w$ indicates C_{max} is the primary objective and F_w is the secondary objective added as constraint.

4. Solution Methodology

Flexible job shop scheduling problems with unrelated parallel machines, sequence and machine-dependent setup times, and/or weighted jobs have been extensively studied in the literature. These problems are notoriously hard to optimize using exact MIP solution approaches, such as Branch-and-Bound or Branch-and-Cut, and the optimality gap of the relaxed problems is traditionally large for instances involving more than 30 to 50 jobs (Rocha et al., 2008). For this reason, a significant body of the literature favors

the development of metaheuristic algorithms (Wilson et al., 2013; Repoussis et al., 2009; Repoussis & Tarantilis, 2010; Tarantilis et al., 2013).

This paper proposes a hybrid multi-start local search framework as shown in Algorithm (1). Initially, a greedy randomized scheme is employed to find and fix part of the patient-to-hospital and patient-to-ambulance assignments (see Line 5). Next, the resulting partially reduced problem is solved to optimality ($\bar{s}$ denotes the partial solution), and a complete solution s is produced by determining the patient dispatching sequence and treatment order at the hospitals (see Line 6). This MIP-based construction heuristic scheme quickly generates a set of high-quality feasible starting upper bounds. Next, these initial heuristic solutions serve as the starting points for an Iterated Tabu Search metaheuristic algorithm, which is mainly applied for further improvement (see Line 7). The oscillations between the MIP-based construction heuristic and the local search algorithm are repeated for a number of iterations ψ_{max} (termination condition). Input parameters ϑ_{max}, λ_{max} and γ control the perturbation mechanism, the maximum local search iterations without observing any improvement, and the randomness of the MIP-based construction heuristic, respectively. Details of these components of Algorithm (1) are provided in the subsections below.

Algorithm 1 Hybrid Multi-Start Local Search

Input: $\psi_{max}, \vartheta_{max}, \lambda_{max}, \gamma, \delta$
Output: s_{best}
1: $\psi \leftarrow 1$, $s_{best} \leftarrow \emptyset$
2: **while** $\psi \leq \psi_{max}$ **do**
3: $s \leftarrow \emptyset$
`// MIP-based Construction Heuristic`
4: $\bar{s} \leftarrow$ Fix Assignments(γ)
5: $s \leftarrow$ MIP Solver($\bar{s}$)
`// Local Search`
6: $s \leftarrow$ Iterated Tabu Search($s, \lambda_{max}, \vartheta_{max}, \delta$)
7: **if** $\{f(s) < f(s_{best})\}$ **or** $\{g(s) < g(s_{best})$ **and** $f(s) <= f(s_{best})\}$ **then**
8: $s_{best} \leftarrow s$
9: **end if**
10: $\psi \leftarrow \psi + 1$
11: **end while**

4.1. MIP-based Construction Heuristic Algorithm

A heuristic decomposition scheme is proposed for generating initial upper bounds. The examined problem involves two sets of decisions, i.e., the assignment of patients to ambulances and hospitals as well as their dispatching and treatment sequencing. On this basis, the proposed MIP-based construction heuristic at first determines the patient-to-ambulance and the patient-to-hospital assignments in a greedy randomized fashion, and subsequently the reduced problem is solved to optimality to obtain the complete solution.

A two-phase sequential assignment scheme is adopted. During the first phase, patients are assigned one by one to hospitals such that the expected completion time of the treatment of each patient is minimized. Based on this greedy criterion, a restricted candidate list of hospitals is generated and maintained for each patient p, and one hospital h_p^* from this list is selected randomly at each iteration. Following a similar greedy randomized assignment scheme, during the second phase the procedure allocates one-by-one each patient p to an ambulance r_p^*, taking into account the previous patients-to-hospitals assignments. Here the greedy criterion is employed to minimize the patient's arrival time at a hospital. The size of the lists in both phases is regulated by parameter γ.

The above heuristic assignments can be depicted as follows:

$$\sum_{b \in B_{h_p^*}} z^b_{ph_p^*} = 1 \qquad \forall p \in P \tag{45}$$

$$\sum_{h \in H \setminus \{h_p^*\}} \sum_{b \in B_h} z^b_{ph} \leq 0 \qquad \forall p \in P \tag{46}$$

$$\sum_{j \in J_{r_p^*}} x^j_{r_p^* p} = 1 \qquad \forall p \in P \tag{47}$$

$$\sum_{r \in R \setminus \{r_p^*\}} \sum_{j \in J_r} x^j_{rp} \leq 0 \qquad \forall p \in P \tag{48}$$

Constraints (45) to (48) are added to the mathematical model that is presented in Section 3.2, and the resulting reduced problem is solved to optimality by a MIP solver. Additionally, an effort is made at the presolve stage to further reduce the number of variables. In particular, given that the number of patients assigned to hospitals and ambulances is fixed, the excessive (if any) positions (beds and trips) in the three-index binary variables z and

x are fix to zero. Note also that all of these variable reductions are not respected during the subsequent local search improvement phase.

4.2. Iterated Tabu Search

The proposed Iterated Tabu Search algorithm has two components; the local search and the perturbation. The local search involves the exploration of the solution space by moving at each iteration from a solution s to the best solution s' of the neighborhood $Nm(s)$, of the neighborhood structure m. Equal selection probability is assumed for all neighborhood structures. The selection of s' follows the hierarchy of objectives defined in Section 3.5. To help the search to escape from local optimal solutions, whenever the local search has performed λ_{max} iterations without observing any improvement, the current best solution is perturbed and the local search restarts.

The solution neighborhoods are created by applying the relocate and exchange operators (Zobolas et al., 2009) on a representation based on the permutation of operations on the machines. A lexicographic scheme is followed for evaluating all allowable combinations for inter- and intra-machine moves. Note that relocations and exchanges can take place only within the same type of the machines, i.e., the ambulances and hospitals separately. Updating the time stamps of this representation is straightforward, but whenever an inter-hospital move is applied the sequence-dependent setup times on the ambulances must be updated accordingly. In our implementation a tabu list is maintained at constant size δ, while the tabu status is overridden if an improvement is observed with respect to the best encountered solution.

As noted, a mechanism is employed to perturb the current solution. In particular, a number of jobs are removed from the schedule and a greedy reconstruction mechanism is employed to reschedule them as early as possible. The number of the rescheduled jobs is determined by a self-adapted "length". The latter is regulated by parameter ϑ. At first, ϑ is initialized to one, and then gradually increases as the search does not find a better solution, until ϑ_{max} is reached. Every time a better solution is found, ϑ is reinitialized to one. At each iteration, $\frac{\vartheta}{\vartheta_{max}}$ % of the solution is perturbed. The above described local search framework is depicted by Algorithm 2.

5. Computational Experiments

We performed a number of computational experiments to study the effectiveness and efficiency of these proposed optimization methods. For this pur-

Algorithm 2 Iterated Tabu Search

Input: $s, \vartheta_{max}, \lambda_{max}, \delta$
Output: s^*

$\vartheta \leftarrow 1, s^* \leftarrow s$
while $\vartheta \leq \vartheta_{max}$ **do**
    $\lambda \leftarrow 1, s' \leftarrow s$
    while $\lambda < \lambda_{max}$ **do**
        $y \leftarrow$ Random Selection()
        $N_y(s) \leftarrow$ Neighborhood Evaluation(s, y)
        $s \leftarrow \min_{s'' \in N_y(s)} \langle f(s''), g(s'') \rangle$
        Update Tabu List(s,y,i,δ)
        if $\{f(s) < f(s')\}$ **or** $\{g(s) < g(s')$ **and** $f(s) <= f(s')\}$ **then**
            $\lambda \leftarrow 1, s' \leftarrow s$
        else
            $\lambda \leftarrow \lambda + 1$
        end if
    end while
    if $\{f(s') < f(s^*)\}$ **or** $\{g(s') < g(s^*)$ **and** $f(s') <= f(s^*)\}$ **then**
        $\vartheta \leftarrow 1, s^* \leftarrow s'$
    else
        $\vartheta \leftarrow \vartheta + 1$
        $s \leftarrow$ Apply Perturbation(s', ϑ)
    end if
end while

pose, we have randomly generated small- and large-scale problem instances based on realistic data. Our methods have been implemented in C++ as sequential programs. All experiments are conducted on a PC equipped with Intel Core i7-4790 clocked at 3.6GHz, 8GB RAM memory and running Windows 7 Professional 64bit edition. IBM Ilog CPLEX 12.60 64bit edition is used as our MIP solver. The default settings are adopted and CPLEX is configured to utilize a single core. A single simulation run is performed for each problem instance, while all computational times reported are in seconds.

This section is structured as follows: In Section 5.1 we describe the data and the parameter settings. Sections 5.2 and 5.3 validate the mathematical model on small-scale instances, and examine the applicability of the model as well as basic problem properties on large scale MCI instances, respectively.

5.1. Experimental Settings & Data Sets

We assume that a hypothetical terrorist bombing attack at the New York Stock Exchange located in the lower Manhattan area leaves a number of critical and non-critical patients in need of emergency medical care. On this basis, a number of problem instances is generated for different MCI events' sizes and resources' availabilities. Table 1 summarizes the main properties of the randomly generated problem instances.

Table 1: Problem instance features and parameter value scenarios

Features	**Scenarios**
Mass casualty event size n_p	Up to 120 patients (large scale) Up to 20 patients (small-scale)
Proportion critically injured	25% - 50% (Baseline: 35%)
Weight w_p for critical patients	10
Treatment time (Non-critical patients)	Normal distribution (40, 10) min
Treatment time (Critical patients)	Normal distribution (120, 25) min
Number of Ambulances n_r	from 3 to 50
Max number of trips	$\frac{n_p}{n_r} + 2$
Number of Hospitals n_h	≤ 4 (small-scale) & ≤ 10 (large scale)
Patient throughput per hospital	1 - 8 per hour

For an event in lower Manhattan, up to 10 hospitals are available to act as local treatment facilities for the patients. Table 2 shows the capacities (in terms of the number of beds), the distance from the disaster site and the patient throughout rates for each hospital. The distances are expressed in terms of traveling time in minutes (see also Section 3.1). The number of beds at each hospital vary from $\frac{n_p}{n_h}$ to n_p, while the patient throughput rates vary from 1-8 patients per hour, based on local knowledge. Note that the hospital beds used in our experiments are only indicative, and they are scaled based on the actual hospital size. Lastly, in every problem instance half of the hospitals are as considered as Trauma Level I and half level II hospitals.

The proposed Hybrid Multi-Start Local Search (HMSLS) uses five parameters. For the experimentation, the following parameter settings are used to perform the experiments. The tabu list size δ is equal to 20. The size of the restricted candidate list γ is equal to 3. The number of tabu search iterations without observing any improvement λ_{max} is set equal to 500, while the maximum number of perturbations ϑ_{max} is equal to 3. Lastly, the number of local search restarts ψ_{max} is equal to 100.

Table 2: Hospital characteristics in Lower Manhattan

Hospitals	Distance (min)	Capacity (beds)	Capacity (max)	Throughput (p/hour)
New York Downtown Hospital	2	n_p/n_h	48	1
Bellevue Hospital Center	11	n_p	370	8
Beth Israel Medical Center	10	n_p	332	6
NY Eye and Ear	9	$n_p/2n_h$	31	1
Hospital For Joint Diseases	10	n_p/n_h	50	2
NY University Medical Center	11	n_p/n_h	212	4
New York Hospital-New York	15	$n_p/2$	644	6
St. Vincents Hospital*	8	$n_p/2$	45	4
St. Lukes Roosevelt/Roosevelt	16	$2n_p/n_h$	323	5
Lenox Hill Hospital	18	$n_p/2n_h$	196	5

*Shuttered after this research was conducted

5.2. Results on Small-Scale Problem Instances

Tables 3 and 4 summarize the results for the small-scale instances with up to 20 patients, 4 hospitals and up to 4 ambulances. Table 3 presents the results derived by running the hierarchical C_{max} as the primary objective, while Table 4 presents the results derived from hierarchical F_w as the primary objective, respectively. In both cases, the constraints on the secondary objectives are derived from the initial MIP start heuristic solutions.

The structure of Tables 3 and 4 is the following. The first two columns report the number of patients and the number of ambulances involved. The third and forth columns report the C_{max} and F_w of the initial upper bounds. These heuristic solutions are generated via the Hybrid Multi-Start Local Search (HMSLS) algorithm and they are used as MIP start solutions. The rest of the columns present the results produced by running CPLEX. A time limit of 7200 sec is assumed. These columns report the C_{max} and F_w as found by CPLEX, the time used to find the best local optimal solution and the number nodes opened during the CPLEX iterations. The last two columns report the optimality % gap between the best upper and lower bound, and the % gap between the solution produced by CPLEX and the heuristic MIP start solution (e.g., $100(C_{max}^{CPLEX} - C_{max}^{Heur.})/C_{max}^{CPLEX}$).

Overall, the following observations can be made. First, it seems very hard to optimally solve instances of more than 12 patients (within the time limit of 2hrs). Second, even though there is sufficient availability of hospital

Table 3: Upper and lower bounds on small-scale instances - Hierarchical $C_{max}|F_w$

Instance		MIP Start (HMSLS)		Best Upper Bound (CPLEX)					
n_p	n_r	C_{max}	F_w	C_{max}	F_w	Time (sec)	Nodes	Gap % (Opt.)	Gap % (Heur.)
8	3	94	2951	94	2951	43.9	27445	0.00	0.00
8	4	93	2772	92	2772	2.2	379	0.00	−1.09
10	3	108	3863	108	3863	1514.7	488021	0.00	0.00
10	4	99	3759	99	3759	0.6	0	0.00	0.00
12	3	125	5043	125	5043	7200	1080612	21.40	0.00
12	4	110	4999	104	4999	85.2	15243	0.00	−5.77
14	3	151	7300	151	7300	7200	421152	32.50	0.00
14	4	123	6268	120	6268	7200	163579	5.30	−2.50
16	3	167	8616	167	8616	7200	388369	29.50	0.00
16	4	148	7333	148	7333	7200	1625623	14.70	0.00
18	3	183	11406	183	11406	7200	258717	22.00	0.00
18	4	159	9837	159	9837	7200	565288	6.00	0.00
20	3	203	12218	203	12218	7200	285077	16.70	0.00
20	4	178	11535	178	11535	7200	404533	2.30	0.00

beds, the response times increase significantly with respect to the number of patients for the same number of ambulances, e.g., the response time doubled from 99 to 178 when the patients double from 10 to 20. Third, the heuristic HMSLS seems to perform reasonably well, while in very few cases CPLEX was able to improve the initial MIP start solutions. Forth, the differences in the solutions between the two tables were relatively small, regardless of the hierarchy of objectives followed. Fifth, when the ambulances were increased from 3 to 4, the optimality gaps were significantly improved. One can lastly observe that the optimality gaps for the hierarchical $C_{max}|F_w$ are better.

5.3. Results on Large Scale Problem Instances

The applicability of the model and the proposed HMSLS has been tested on large scale problem instances with up to 150 patients. In particular, various scenarios were examined using different number of ambulances and hospitals. The goal is to answer the following realistic logistical question for emergency service providers: is it better to send patients to remotely located

Table 4: Upper and lower bounds on small-scale instances - Hierarchical $F_w|C_{max}$

Instance		MIP Start (HMSLS)		Best Upper Bound (CPLEX)					
n_p	n_r	C_{max}	F_w	C_{max}	F_w	Time (sec)	Nodes	Gap % (Opt.)	Gap % (Heur.)
8	3	98	2823	98	2811	174.4	67126	0	−0.43
8	4	93	2772	93	2748	70.8	31465	0	−0.87
10	3	124	3801	124	3741	7200	1665570	2.7	−1.60
10	4	109	3703	109	3668	3532.4	543242	0	−0.95
12	3	140	4919	140	4919	7200	186839	29.1	0.00
12	4	127	4738	127	4738	7200	131585	26.5	0.00
14	3	156	6095	156	6095	7200	461692	32.9	0.00
14	4	144	5899	144	5766	7200	1179001	21.7	−2.31
16	3	179	7521	179	7521	7200	431088	51.1	0.00
16	4	155	7281	155	7281	7200	82222	32.2	0.00
18	3	200	9513	200	9513	7200	172898	94.7	0.00
18	4	173	9183	173	9183	7200	333900	58.9	0.00
20	3	242	11599	242	11599	7200	144523	115.4	0.00
20	4	190	11157	190	11157	7200	37435	99.4	0.00

hospitals when optimizing response for a large MCI?

Table 5 summarizes the results obtained by applying HMSLS on large scale problem instances for the hierarchical $C_{max}|F_w$ and $F_w|C_{max}$. The first four columns refer to the problem instance, showing the number of patients, the number of ambulances, and the number of hospitals, respectively. Columns C_{max} and F_w report the makespan and the weighted total flow time. Columns TD report the total distance traveled by the ambulances (including so-called deadhead trips, in which an ambulance travels back to the MCI from drop off without any passengers); the lower bound for this distance is shown in column LB^d. The latter is calculated by dispatching the ambulances to the closest hospitals, while taking into account the maximum number of beds per hospital as a constraint. Lastly, column LB^t reports the lower bound of the makespan (see Equation 40).

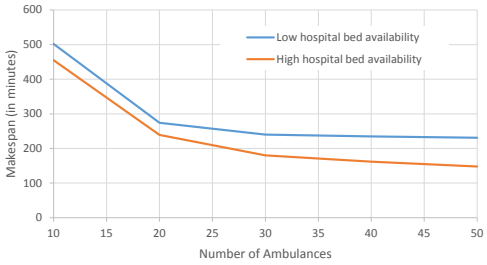
Table 5 clearly demonstrates that the response times are affected by both the available ambulances and hospitals. As the number of hospitals increases, the response time is largely reduced. For example, regarding the problem

Table 5: Upper bounds on large scale instances

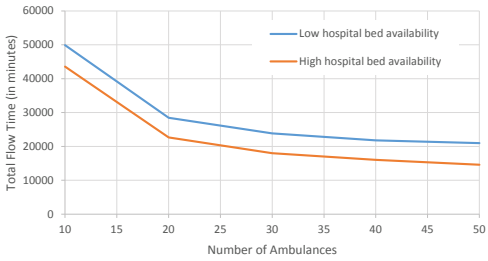
Instances						HMSLS - $C_{max}\|F_w$			HMSLS - $F_w\|C_{max}$		
♯	n_p	n_r	n_h	LB^t	LB^d	C_{max}	F_w	TD	C_{max}	F_w	TD
1	30	20	6	68.67	258	196	19781	688	206	19220	691
2	30	20	8	52.50	245	122	13534	642	128	13524	801
3	30	20	10	42.80	245	107	12359	791	108	11788	844
4	50	30	6	116.00	399	315	48311	1183	326	47875	1202
5	50	30	8	87.12	408	188	31550	1274	196	31062	1373
6	50	30	10	67.00	432	153	26864	1409	149	24541	1491
7	70	30	6	162.00	607	441	91392	1789	445	90194	1694
8	70	30	8	122.12	571	260	58670	1937	274	56639	2019
9	70	30	10	95.30	562	206	47165	2041	208	44091	2163
10	70	40	6	164.00	607	436	93479	1771	452	90391	1900
11	70	40	8	120.12	571	257	57386	1928	274	55274	1947
12	70	40	10	93.30	562	207	48695	1928	197	42419	2084
13	90	40	6	208.50	773	562	147718	2239	567	144520	2195
14	90	40	8	153.00	734	333	92115	2335	343	91445	2494
15	90	40	10	122.30	717	266	75955	2606	272	73481	2785
16	90	50	6	210.50	773	559	149254	2379	549	145657	2317
17	90	50	8	155.00	734	329	92308	2478	361	87641	2510
18	90	50	10	120.30	717	260	74864	2566	249	68068	2713
19	110	30	8	189.60	897	407	134097	3190	410	136205	3210
20	110	30	10	149.00	880	327	110945	3410	330	107199	3437
21	110	40	8	187.60	897	411	136937	3049	404	133716	3047
22	110	40	10	147.00	880	324	109981	3285	316	106925	3352
23	110	50	8	187.60	897	399	132118	2942	410	130286	2963
24	110	50	10	147.50	880	316	108379	3141	298	96529	3421
25	130	30	8	223.25	1060	479	190292	3823	484	183498	3820
26	130	30	10	176.00	1034	390	154140	4136	405	149564	4085
27	130	40	8	221.50	1060	479	185573	3715	459	173303	3800
28	130	40	10	174.00	1034	384	154924	3994	385	148472	4061
29	130	50	8	219.25	1060	470	182563	3558	480	179898	3636
30	130	50	10	172.00	1034	378	149897	3902	381	146326	3840
31	150	30	8	260.37	1223	557	252585	4435	569	247937	4452
32	150	30	10	204.50	1197	445	201296	4801	464	199081	4770
33	150	40	8	256.37	1223	553	246023	4338	540	234263	4404
34	150	40	10	200.50	1197	439	198331	4646	443	194689	4644
35	150	50	8	256.37	1223	536	242509	4223	543	237009	4287
36	150	50	10	200.50	1197	430	196787	4549	432	180210	4702

instances with 30 patients (see Pr.No 1 and 3) increasing hospitals from 6 to 10 reduced the response time from 196 to 107, while the addition of 2 hospitals for instances Pr.No 31 and Pr.No 32 with 150 patients reduced the response time from 557 to 445. The figures with respect to the number of ambulances shows similar improvement, although the effect is smaller. Note that in several cases, while the improvements in the primary objective are small, the reductions in the secondary objective are more substantial (see for example Pr. No 25, 27 and 29). In answer to our question regarding the role of distant hospitals, we find that adding remotely located facilities increases the traveled distance but reduces the makespan (see for example Pr.No. 16 to 18). This is an indication that although longer transportation times are introduced, the allocation of patients to hospitals is more balanced and the capacity of the hospital system taken as a whole is more effectively utilized. Lastly, it seems that in most cases the traveled distance by the ambulances (TD) is lower when the primary objective is to minimize the makespan.

We next examined in more detail the effect of the number of ambulances on the response time. In particular, we consider an MCI event with 120 patients, high and low bed availability at 10 hospitals, and variation in the number of ambulances from 10 to 50. Figure 2 illustrates the results obtained from these experiments. Specifically, Figure 2 (a) refers to the hierarchical $C_{max}|F_w$ and Figure 2 (b) shows the hierarchical $F_w|C_{max}$.



(a) Makespan $C_{max}|F_w$ vs No of ambulances



(b) Total flow time $F_w|C_{max}$ vs No of ambulances

Figure 2: The effect of the number of ambulances to the response time

Overall, three main observations can be made. First, as the number of ambulances increase the response times improve; however, this effect quickly fades out. This is likely due to the fact that the hospitals have become the bottleneck. Second, the high or low availability of beds at the hospitals has a relatively small impact on overall outcomes. This is reflected in practice,

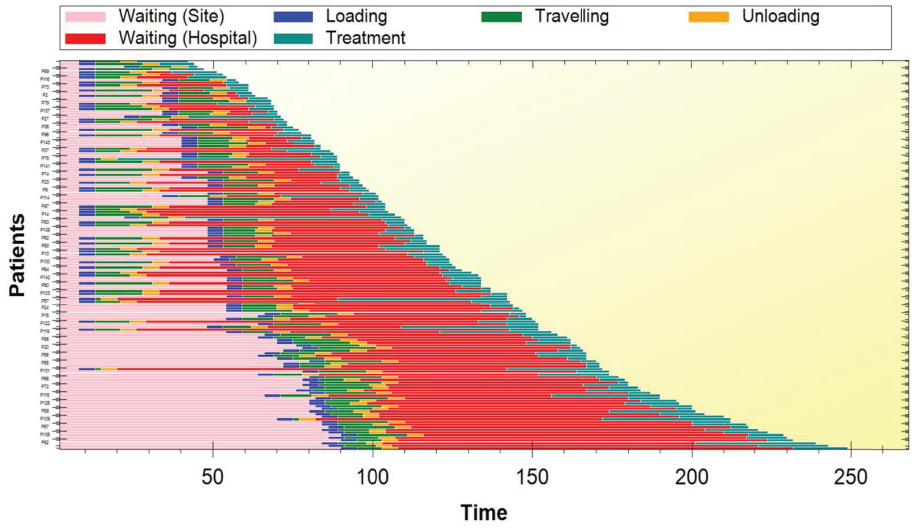
in which inquiries are mostly made regarding the availability of emergency department beds, and not regarding hospital census on the inpatient wards. However, this effect tends to be stronger when the hospital beds become the bottleneck (case with 50 ambulances), a finding that may have important policy implications regarding current MCI dispatching practices. Lastly, the hierarchy of objectives has little differential impact on the observed trends.

Figure 3 summarizes our findings graphically by presenting Gantt charts of two indicative schedules for a response to a large-scale MCI with 150 patients, produced during the execution of the HMSLS. The first solution (Figure 3 (a)) corresponds to the best solution found by the MIP-based construction heuristic, while the second (Figure 3 (b)) corresponds to the overall best solution found, having also applied the Iterated Tabu Search algorithm. As Figure 3 shows, different colors represent different activities throughout the schedule of each patient (see also timing of events Figure 1). One can observe that the Iterated Tabu Search significantly improves the initial MIP based construction heuristic solution, indicating the effectiveness of the proposed local search improvement method; this is also consistent in all computational experiments performed. Second, the waiting times at hospitals (red color) comprises a major part of the schedule of virtually all patients; this is a vivid indication of the important influence of treatment sequencing at the hospitals on overall outcomes, which makes the combined ambulance dispatching-treatment ordering problem very hard to solve.

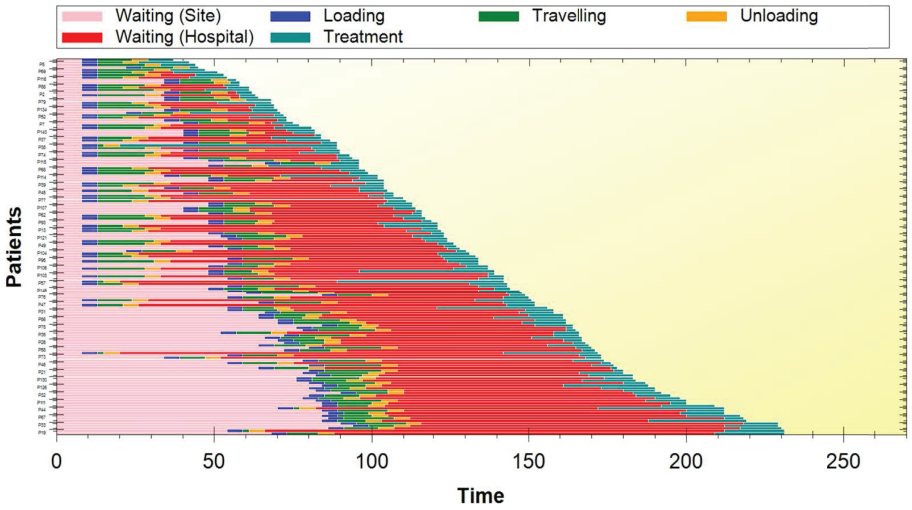
6. Conclusions

This paper linked pre-hospital emergency medical services with hospital treatment activities to advance the science of disaster preparedness planning, presenting a rigorous MIP model for the combined ambulance dispatching, patient-to-hospital assignment, and treatment ordering problem. The objective was to minimize the makespan and the weighted total flow time of the patients hierarchically. Both exact and hybrid metaheuristic methods were developed for small- and large-scale problems. Regarding the latter, a heuristic problem decomposition scheme is adopted that utilizes a MIP-based construction heuristic combined with an Iterated Tabu Search algorithm.

We performed various experiments to establish the validity of the model in relation to known features of MCI response and to assess the performance of the proposed methods. Furthermore, we examined how the spatial and temporal characteristics of an MCI response–especially the inclusion of re-



(a) MIP-based construction heuristic solution



(b) Iterated Tabu Search solution

Figure 3: Schedule of an MCI event with 150 patients

mote hospitals for large events–affect the response times and the system-wide allocation of resources. Finally, we demonstrated the applicability of the new model on MCI events, showing how response efficiency is impacted by varying resource availability at the ambulance and hospital levels.

We believe that the proposed model has practical utility in helping emergency response professionals to explore tactical decision making for disaster preparedness, specifically for events with many casualties. Astute readers will be aware of several limitations in our formulation, including the assumption that the patient load does not impact the functioning of near hospitals (as may be the case in a large-scale explosive or contamination event). This is related to our assumption that there is considerable information availability regarding hospital capacities and capabilities, which may change early on ("fog of war" effect). It will be of great value to extend the current model for stochastic environments with incomplete information regarding treatment capability. Finally, this modeling approach provides a sophisticated platform to explore advanced topics, such the capacity-limiting impact of overtriage and of self-transported patients on the overall response effectiveness.

Acknowledgements

Dr. Nathaniel Hupert receives support from the Cornell Institute for Disease and Disaster Preparedness, which is funded New York-Presbyterian Hospital and Weill Cornell Medicine. We would like to thank Horia Tipi, Gabriel Tavares and Richard Laundy (now FICO/Fair Isaac previously Dash Optimization) for their effort in the initial work of this research.

References

Amram, O., Schuurman, N., & Hedley, N. (2012). A web-based model to support patient-to-hospital allocation in mass casualty incidents. *The Journal of Trauma*, *72*, 1323–1328.

Arnold, J., Halpern, P., Tsai, M., & Smithline, H. (2004). Mass casualty terrorist bombings: a comparison of outcomes by bombing type. *Annals of Emergency Medicine*, *43*, 263–273.

Bae, J., Shin, K., Lee, H., Lee, H., Lee, T., Kim, J., Cha, W., Kim, G., & Moon, I.-C. (2015). *Evaluation of disaster response system using agent-based model with geospatial and medical details*. Working Paper Applied Artificial Intelligence Laboratory, Korea.

Barbarosoglu, G., & Arda, Y. (2004). A two-stage stochastic programming framework for transportation planning in disaster response. *Journal of the Operational Research Society*, *55*, 43–53.

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

Bektas, T., Repoussis, P., & Tarantilis, C. (2014). Dynamic vehicle routing problems. In P. Toth, & D. Vigo (Eds.), *The Vehicle Routing Problem* Monographs on Discrete Mathematics and Optimization chapter 11. (pp. 299–347). SIAM, Philadelphia, US. (2nd ed.).

Boldberg, J. (2004). Operations research models for the deployment of emergency service vehicles. *EMS Management Journal*, *1*, 20–39.

Brotcorne, L., Laporte, G., & Semet, F. (2003). Ambulance location and relocation models. *European Journal of Operations Research*, *147*, 451–463.

Budge, S., Ingolfsson, A., & Zerom, D. (2010). Empirical analysis of ambulance travel times: the case of calgary emergency medical services. *Management Science*, *56*, 716–723.

Carter, G., Chaiken, J., & Ignall, E. (1972). Response areas for two emergency unites. *Operations Research*, *20*, 571–594.

Chiu, Y.-C., & Zheng, H. (2007). Real-time mobilization decisions for multi-priority emergency response resources and evacuation groups: Model formulation and solution. *Transportation Research Part E: Logistics and Transportation Review*, *43*, 710–736.

Dean, M., & Nair, S. (2014). Mass-casualty triage: distribution of victims to multiple hospitals using the save model. *European Journal of Operational Research*, *238*, 363–373.

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

Fitzsimmons, J. (1973). A methodology for emergency ambulance deployment. *Management Science*, *19*, 627–636.

Gendreau, M., Laporte, G., & Semet, F. (2006). The maximal expected coverage relocation problem for emergency vehicles. *Journal of the Operational Research Society*, *57*, 22–28.

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

Haghani, A., & Yang, S. (2007). Real-time emergency fleet deployment. In V. Zeimpekis, C. Tarantilis, G. Giaglis, & I. Minis (Eds.), *Dynamic Fleet Management: Concepts, Systems, Algorithms and Case Studies* (pp. 133–162). Springer, New York, USA volume 38.

Hanfling, D., Altevogt, B., Viswanathan, K., & Gostin, L. (2012). *Committee on guidance for establishing crisis standards of care for use in disaster situations*. Technical Report Institute of Medicine.

der Heide, E. A. (2006). The importance of evidence-based disaster planning. *Annals of Emergency Medicine*, *47*, 34–49.

Hoogeveen, H. (2005). Multicriteria scheduling. *European Journal of Operational Research*, *167*, 592–623.

Hupert, N., Hollingsworth, E., & Xiong, W. (2007). Overtriage and outcomes: insights from a computer model of trauma system mass casualty response. *Disaster Medicine and Public Health Preparedness*, *1*, 14–24.

Jotshi, A., Gong, Q., & Batta, R. (2009). Dispatching and routing of emergency vehicles in disaster mitigation using data fusion. *Socio-Economic Planning Sciences*, *43*, 1–24.

Larson, R. (1974). A hypercube queueing model for facility location and redistricting in urban facility service. *Computers & Operations Research*, *1*, 67–95.

Mete, H., & Zabinsky, Z. (2010). Stochastic optimization of medical supply location and distribution in disaster management. *International Journal of Production Economics*, *126*, 76–84.

Mills, A., Argon, N., & Ziya, S. (2014). Resource-based patient prioritization in mass-casualty incidents. *Manufacturing and Service Operations Management*, *15*, 361–377.

Najafi, M., Eshghi, K., & Dullaert, W. (2013). A multi-objective robust optimization model for logistics planning in the earthquake response phase. *Transportation Research Part E Logistics and Transportation Review*, *49*, 217–249.

Najafi, M., Eshghi, K., & de Leeuw, S. (2014). A dynamic dispatching and routing model to plan/re-plan logistics activities in response to an earthquake. *OR Spectrum*, *36*, 323–356.

Ozdamar, L. (2011). Planning helicopter logistics in disaster relief. *OR Spectrum*, *33*, 655–672.

Park, J., Shin, S., Song, K., Hong, K., & Kim, J. (2016). Epidemiology of emergency medical services-assessed mass casualty incidents according to causes. *Journal of Korean Medical Science*, *31*, 449–456.

Repoussis, P., Paraskevopoulos, D., Vazacopoulos, A., & Hupert, N. (2015). Optimizing emergency preparedness and resource utilization in mass-casualty incidents. In *INFROMS Business Analytics and Operations Research Conference 2015*. Los Angeles, California, USA.

Repoussis, P., & Tarantilis, C. (2010). Solving the fleet size and mix vehicle routing problem with time windows via adaptive memory programming. *Transportation Research Part C*, *18*, 695–712.

Repoussis, P., Tarantilis, C., & Ioannou, G. (2009). Arc-guided evolutionary algorithm for the vehicle routing problem with time windows. *IEEE Transactions on Evolutionary Computation*, *13*, 624–647.

Rocha, P., Ravetti, M., Mateus, G., & Pardalos, P. (2008). Exact algorithms for a scheduling problem with unrelated parallel machines and sequence and machine-dependent setup times. *Computers & Operations Research*, *35*, 1250–1264.

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

Salman, F., & Gul, S. (2014). Deployment of field hospitals in mass casualty incidents. *Computers & Industrial Engineering*, *74*, 37–51.

Sung, I., & Lee, T. (2016). Optimal allocation of emergency medical resources in a mass casualty incident: Patient prioritization by column generation. *European Journal of Operational Research*, *252*, 623–634.

Tarantilis, C., Anagnostopoulou, A., & Repoussis, P. (2013). Adaptive path relinking for vehicle routing and scheduling problems with product returns. *Transportation Science*, *47*, 356–379.

Toregas, C., Swain, R., ReVelle, C., & Bergman, L. (1971). The location of emergency service facilities. *Operations Research*, *19*, 1363–1373.

Wilson, D., Hawe, G., Coates, G., & Crouch, R. (2013). A multi-objective combinatorial model of casualty processing in major incident response. *European Journal of Operational Research*, *230*, 643–655.

Wilson, D., Hawe, G., Coates, G., & Crouch, R. (2016). Online optimization of casualty processing in major incident response: An experimental analysis. *European Journal of Operational Research*, *252*, 334–348.

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

Yi, W., & Ozdamar, L. (2007). A dynamic logistics coordination model for evacuation and support in disaster response activities. *European Journal of Operational Research*, *179*, 1177–1193.

Zobolas, G., Tarantilis, C., & G.Ioannou (2009). A hybrid evolutionary algorithm for the job shop scheduling problem. *Journal of the Operational Research Society*, *60*, 221–235.

Zoraster, R., Chidester, C., & Koenig, W. (2007). Field triage and patient maldistribution in a mass-casualty incident. *Prehospital and Disaster Medicine*, *22*, 224–229.