Decision Support for Casualty Triage in Emergency Response

Behrooz Kamali

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Industrial and Systems Engineering

Douglas R. Bish, Chair
Ebru K. Bish
Roger E. Glick
G. Don Taylor
Christopher W. Zobel

March 25, 2016
Blacksburg, Virginia

Keywords: Emergency Management, Operations Research, Mass Casualty Triage

Decision Support for Casualty Triage in Emergency Response

Behrooz Kamali

Abstract

Mass-casualty incidents (MCI) cause a sudden increase in demand of medical resources in a region. The most important and challenging task in addressing an MCI is managing overwhelmed resources with the goal of increasing total number of survivors. Currently, most of the decisions following an MCI are made in an ad-hoc manner or by following static guidelines that do not account for amount of available resources and number of the casualties. The purpose of this dissertation is to introduce and analyze sophisticated service prioritization and resource allocation tools. These tools can be used to produce service order strategies that increase the overall number of survivors. There are several models proposed that account for number and mix of the casualties, and amount and type of the resources available. Large number of the elements involved in this problem makes the model very complex, and thus, in order to gain some insights into the structure of the optimal solutions, some of the proposed models are developed under simplifying assumptions. These assumptions include limitations on the number of casualty types, handling of deaths, servers, and types of resources. Under these assumptions several characteristics of the optimal policies are identified, and optimal algorithms for various scenarios are developed. We also develop an integrated model that addresses service order, transportation, and hospital selection. A comprehensive set of computational results and comparison with the related works in the literature are provided in order to demonstrate the efficacy of the proposed methodologies.

This work was supported in parts by the National Science Foundation (Grant #1055360) and Carilion Clinic.

Dedication

To my lovely wife, Leily.

Acknowledgments

I would like to extend special thanks to my advisor, Douglas Bish, and the rest of my committee members, Ebru Bish, Roger Glick, Don Taylor, and Christopher Zobel. I would also like to thank the faculty members and staff at the Grado Department of Industrial and Systems Engineering for their support. Finally, I would like to thank Leily Farrokhvar, my best friend and partner in life, to have always been my inspiration. I owe you a lot, love you forever.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Mass casualty incidents (MCI) involve large number of injured people compared to the level of regional resources available. MCIs are rare, and their infrequent nature and large impact make effective preparation and proficient management challenging. On the other hand, the frequency and scale of MCIs, both natural and man-made, has been increasing significantly in the past few decades. This is in part due to population growth and spread, urbanization, development of advanced technologies, hazardous material, economic imbalance, and rise of infectious diseases (Arnold, 2002). Sudden increase in demand for service to casualties overwhelms the medical resources (e.g., hospital beds, ambulances, physicians, nurses, operating rooms, supplies, etc.). As a result, resources should be rationed among casualties to ensure those in a higher need are prioritized. In many cases, there are multiple care providers in the affected region. While this can be an opportunity to serve a larger number of casualties in a timely manner, it brings several challenges for management of resources in the short window of time available. These challenges highlight the need for effective preparedness and efficient management of MCIs to allocate scarce resources in order to "do the greatest good for the greatest number" (Antommaria et al., 2011).

In the aftermath of an MCI, the first step upon arrival of medical personnel is to triage the casualties. Triage refers to the categorization of casualties based on the severity their con-

dition (e.g., Simple Triage and Rapid Treatment (START), Homebush, Triage Sieve, Sacco Triage Method (STM), CESIRA, NATO Triage, and etc. (Lerner et al., 2008)). Although there are several triage methods, most of them have several characteristics in common. For instance, most triage methods use a "walking filter" to identify casualties with minor injuries quickly and postpone service to them. Casualties with no chance of survival are often categorized as "expectant" or "morgue". To ease the identification, categories are color coded, with most triage methods having red for most critical, yellow for critical, green for minor, and black for deceased casualties. The main difference between triage methods is the criteria and methods based on which casualties are categorized. Lerner et al. (2015a) identify the lack of standard definition for casualty categories as the main obstacle in comparing the performance between triage systems and develop consensus-based definitions for each MCI triage category. With improvements in data collection during MCIs in the past decades, Wolf et al. (2014) develop ASAV (Amberg-Schwandorf Algorithm for Primary Triage) triage method with categorization derived based on the data from survival of casualties in the past incidents. Next step after categorization of casualties is prioritization for receiving care or transportation. No triage method is confirmed to perform better than any other in terms of survived casualties, scene management, or resource allocation (Kahn et al., 2010).

START, which is the most common triage method in the US, uses several physiological parameters including respiration, pulse rate, and mental status with the goal of completing casualty's assessment in less than 60 seconds. It has four categories; Immediate (red), delayed (yellow), minor (green), and deceased (black). Following assessment, each casualty is tagged with their color-coded category, ready for treatment or transportation. Homebush triage standard is based on START, with minor differences in categorization process. In addition, it also has a dying category (white) to separate those who are dying from already deceased. Triage Sieve is similar to START with four categories. Sacco triage methods calculates a score from 0 to 12 for each casualty using a computer software based on historical data (Sacco et al., 2005, 2007). CESIRA has three categories, dropping the deceased category. The red category is for casualties that are unconscious, hemorrhaging, in shock, or with insufficient

respiration, yellow category for broken bones and other injuries, and green category for walking. Most triage methods come with an inherent order for service, which is from most critical (red), to less critical (yellow), and finally to minor (green) casualties. One exception is Sacco triage method, which calculates the service order using their computer software.

1.1 Shortcomings of Current Triage Methods

There has not been many studies evaluating performance of triage methods or precisely documenting how they performed in past incidents. Kahn et al. (2010) mention that the use of START has been documented in two terrorist attacks and two natural disasters in the US in the past two decades. However, those studies are descriptive and provide no detailed data about casualties and survivors. Several lines of work in the literature have raised the need for scientific evaluation of effectiveness and efficiency of START and other triage methods (Jenkins et al., 2008; Lerner et al., 2008). In addition, current triage methods have multiple shortcomings, which directly affect their performance. Triage methods have a static nature and they treat various situations in the same manner and do not account for disaster-specific characters (Kienstra and Endom, 2002). It is not possible to train and prepare for every single type of disaster, instead, there should be policies to aid management of various scenarios to maximize expected outcome.

The majority of triage processes do not consider the level of medical resources in relation to the number of casualties or the mix of casualty types when making decisions. One study incorporates resource utilization in categorizing burn casualties (Taylor et al., 2014), but as mentioned earlier, most triage methods come with an inherent order of service, which is providing care to casualties from most critical to least critical. After categorization of casualties, they will be transported to nearby hospitals, either based on the mentioned static ordering policy or in an ad-hoc manner. The main goal of triage is to prioritize already overwhelmed resources such that expected number of the survivors is maximized. This

cannot be achieved without consideration of the resources in the process. There are a few studies that look into how consideration of the resources in the triage process affects the outcomes under simplifying assumptions. Jacobson et al. (2012) develop a model for priority assignment of two casualty types with one server, under the assumptions of exponentially-distributed lifetimes and that only casualties who are going to survive, are served, and thus, no casualty dies after service. They also develop several simple heuristics and compare their performance. They analyze some special cases of the problem including having multiple identical hospitals and Weibull-distributed lifetimes. Mills et al. (2013) develop a fluid formulation for the two-casualty types case, under the assumption that no casualty dies prior to receiving service. They also develop two heuristics, one dynamic and one static, and study cases with multiple identical emergency vehicles for transferring the casualties to a care provider. In another study, Sacco et al. (2005) study combination of service order and resource allocation, in which only transportation resources are studied in a high level. Under simplifying assumptions, these studies show that considering resources could potentially improve the triage process.

Following categorization and prioritization of casualties, they need to be assigned to an emergency vehicle and a receiving hospital based on availability. Currently, to the best of our knowledge, these decisions are made in an ad-hoc manner. Care capability and service level among different hospitals make choosing the receiving hospital one of the most challenging decision (Schneider et al., 2003). There are typically multiple healthcare facilities available in a region. Arbitrary assignment of casualties to nearby hospitals leads to inefficient utilization of beds. Overcrowding, in addition to decrease in level of care, causes financial loss, impaired access, and in the worst case mortality (Hoot and Aronsky, 2008). Bagust et al. (1999) show through a series of simulation studies that risk of poor outcome increases as bed occupancy increases in a hospital. The decrease is sharper as occupancy rises above 95%. There is also a direct relationship between hospital and emergency department overcrowding and death of casualties (Sprivulis et al., 2006). Thus, it is critical for emergency personnel to assign casualties to healthcare providers efficiently in order to maintain quality of care and the

lower risk of adverse outcomes.

Apart from the service order, there are several other decisions that emergency personnel need to make following an MCI, including vehicle assignment, receiving hospital selection, and allocation of the resources. Current triage methods either do not provide any framework for making such decisions or provide static policies, thus, they are made by physicians and emergency personnel in an ad-hoc manner. One example for this happened during the Hurricane Katrina in August 2005. Many physicians who were in charge, did not have prior training in triage and some of the made decisions were emotional rather than logical (Darr, 2006). Another challenge is when multiple organizations are collaborating to provide service to the same MCI, and they might use different triage methods (Lerner et al., 2015b). In addition to the triage process itself, related decisions such as assigning and staffing ambulances could become challenging when multiple organizations are involved in providing care to MCI casualties (Griffiths et al., 2014). This also highlights another importance of having a comprehensive triage and resource allocation framework to generate effective and efficient policies under a wide range of scenarios.

1.2 Capacity Management

Operations prior to transferring casualties are critical to ensure timely service, but due to large number of casualties, it is most likely that there is not enough beds and resources to serve casualties in healthcare facilities. Management of casualties require efficient allotment of limited resources to cope with the surge in demand. Internal operations under catastrophic circumstances require specific attention due to their difference from daily operations. The main differences are how operations are extended and standards of care are altered. Surge capacity planning and management was brought into attention in the past several years, following several disasters such as September 11, 2001 terrorist attacks and Hurricane Katrina. Medical processes involved in providing care to casualties affected by disasters have

been studied in details, however, administration of these tasks and procedures has received insufficient attention.

There are several definitions for surge capacity, but none is commonly accepted among scholars and practitioners (Hick et al., 2009). Nager and Khanna (2009) define surge capacity as "health care systems' ability to rapidly expand normal services to meet the increased demand for qualified personnel, medical care, and public health, in the event of bio-terrorism or other large-scale public health emergencies or disasters". Among the existing definitions of surge capacity, to our knowledge, only one focuses on management aspect; Surge capacity translates to the maximum potential delivery of required resources, either through expansion or alteration of resource management and allocation (Kelen and McCarthy, 2006). In fact, some scholars highlight what surge capacity planning is lacking, is the right management, not the personnel and resources. Management in surge capacity planning can be interpreted as applying the right resources to the right place at the right time (Koenig et al., 2006). Main elements of surge capacity management are identified as: 1) system (structure), 2) staff, 3) stuff (supplies), and 4) space (Nager and Khanna, 2009). System is the most important element, as performance of other elements is directly affected by it.

There are two main groups of actions identified in the literature to increase the surge capacity; altering standards of care and extending operations. Standard of care is defined as "the level at which the average, prudent provider in a community would practice" (Koenig et al., 2006). There are several alterations to the standards of care suggested in the literature including early discharge of inpatients, transfer of less-critical casualties to alternate care facilities such as nursing homes, cancellation of nonessential surgeries, medication substitutions, and shelf life extension.

In addition to mentioned guidelines for altering standards of care and extension of operations, there are several high level frameworks proposed for surge capacity management. Most frameworks divide service into multiple stages, starting from local and extending to national level. Bonnett et al. (2007) developed a surge capacity response framework consisted of four

stages. Stage 1 is daily operations level, if regular staff and resources could address the demand. If resources are insufficient, stage 2 is to extend operations and alter standards of care. Subsequently, stage 3 is to extend operations beyond the hospital and use alternate care facilities, and finally if demand still cannot be met, in stage 4 outside help such as federal support should be sought. There are other frameworks such as 6-tiered response system (Barbera and Macintyre, 2007) and 3 category framework called CO-S-TR (Hick et al., 2008). Majority of these frameworks act as guidelines on how to extend operations without consideration of characteristics of the disaster. A look at all these frameworks and guidelines shows that they all try to improve the operations and increase the surge capacity in some way, but they do not precisely study how operations leading to increase in surge capacity should be managed. In addition, these frameworks all have a subjective nature with no measurable criteria for level of care or precise details on how to move from one tier to another. While mentioned frameworks are developed to address a wide range of population types, due to different characteristics of pediatric casualties, they need specific attention that is discussed in the next section.

1.3 Pediatric Casualties

Anatomical and physiological differences between pediatric and adult casualties can make children more vulnerable to disasters (Gausche-Hill, 2009). For instance, children have smaller volume of circulating blood that makes them more exposed to loss of fluids. Also, due to smaller body area, same amount of force could cause more damage to multiple organs compared to adults. Most triage methods lack consideration for population-specific needs such as neonatal and pediatric casualties (Barfield et al., 2011). There are two pediatric-specific triage methods; JumpSTART and Pediatric Triage Tape (PTT). JumpSTART is based on START, with minor modifications to account for anatomical differences of children and PTT is based on Triage Sieve. Jones et al. (2014) compare the performance of JumpSTART with SALT (Sort, Assess, Lifesaving interventions, Treat/Transport), which is proposed for both

adults and pediatrics. They find that SALT is at least as good as JumpSTART in overall triage accuracy, but JumpSTART is faster around 8 seconds per casualty.

Due to smaller population of children, their better health condition, and higher resiliency compared to adults, there are less children hospitals and pediatric-specific units in adult hospitals. In fact, only 5.5% of US hospital emergency departments have all the recommended pediatric equipment (Lyle et al., 2009). This hurts surge capacity management for pediatric patients. In an empirical study of hospitals in the New York state, it was found that while the surge capacity suffice for adults, there are not enough beds available for pediatric patients (Kanter and Moran, 2007a). In another study, a simulation analysis is used to show how altering standards of care can increase surge capacity for pediatric casualties (Kanter and Moran, 2007b). They show that when all resources are available, 250 pediatric casualties per million population can be served under regular standards of care in New York city, while under altered standards of care with 40% of resources available, up to 500 pediatrics can be served, which is still below the target capacity in most categories. Sometimes lower capacity for pediatric casualties forces emergency personnel to transfer a portion of them to adult care facilities. In addition to considering risk exposed by surge capacity in care facilities, specialized resources and capabilities should be taken into account when deciding about receiving hospital for pediatrics. Toltzis et al. (2015) suggest categorizing pediatric casualties based on whether they need mechanical ventilation at a pediatric intensive care unit (PICU) or not, and then, estimating a probability of death and duration of resource consumption for each casualty requiring PICU. Depending on number of the casualties and amount of resources available, low and high risk thresholds are defined, and casualties within this range are considered for admission to available PICUs. Although prioritization is not directly considered in this study, resource availability is included in the categorization of the casualties. Toltzis et al. (2013) also develop a resource allocation scheme for pediatric casualties during an MCI, with the main goal of assigning limited number of PICUs to the casualties with a higher chance of survival.

There are several other challenges regarding providing care to pediatric casualties. A major

concern is separation of children from their parents (Lyle et al., 2009). In the aftermath of Katrina and Rita, 5,000 children were reported missing and it took 6 month for the last children to be reunited with their family (Gausche-Hill, 2009). Communication could also be nonexistent to limited with neonatal and younger pediatric casualties, which slows down relief efforts. While adult casualties can communicate their condition with emergency personnel, in many cases pediatrics cannot and personnel have to rely on their tests and opinions.

1.4 Overview

The work in this dissertation is divided into three papers. For the first paper, processes in the aftermath of a mass-casualty incident (MCI) are studied. One of the first steps in the response is to triage the casualties. Triage systems categorize the casualties into casualty types based on their criticality, and then prioritize casualties for transfer to a hospital for further treatment. The prioritization is usually based on simply ordering the casualty types without considering the available resources to transport them and the scale of the disaster. These factors can significantly affect the outcome of the rescue efforts. In this paper we develop a mathematical model to incorporate the above mentioned factors in the triage process. It is assumed that there is a disaster location with a set of casualties, categorized by criticality and care requirements. These casualties need to be transported to hospitals in the region. There is a limited fleet of vehicles available to transfer the casualties. The goal of the developed model in this paper is to maximize the expected number of survivors. We analyze the structure of the optimal solution to this problem, and compare the performance of our model with the current practice and other related models in the literature.

In the second paper, the focus is on different approaches to model the prioritization in the triage process. To the best of our knowledge, all the related papers in the literature either assume that all the MCI casualties are served with no deaths observed prior to service (on-

site death), or if an on-site death occurs, they will capture that immediately and remove the deceased casualty from the system with no service. On-site death refers to the degradation in a casualty's condition due to delay in service to a point that their survival probability is zero. In the former case, the assumption is that there is no information regarding casualty's death and therefore, all casualties are served. Service to a deceased casualty (i.e., a casualty with no chance of survival regardless of the available resources) refers to occupying resources without receiving any reward. In the latter case, we assume there is perfect information about casualty's death, and once a death occurs, that casualty is removed from the system without receiving service. While these simplifications allow for a more detailed analysis of the problem and structure of the optimal policies, in practice a combination of these cases could occur. We model the general problem and compare its results to that of the simplified cases. We analyze how the data used in the problem could affect the results, and suggest a practical set of scenarios for testing purposes. Based on the structural analysis of the optimal policies, we propose a heuristic approach which solves the problem to optimality under the case in which all casualties are served, and provides high quality results under other scenarios.

For the last paper we are extending the domain of the study to include casualty transportation and hospital assignment. While it is important to analyze parts of the triage process separately to gain some insight by reducing the complexity, this problem is a complicated problem by nature with many element interacting. MCIs overwhelm all medical resources in a region including transportation resources and hospital beds and staff. We develop a mixed-integer programming model with the goal of studying how resource constraints in hospitals affect the results and optimal policies in triage operations. To further account for the mentioned elements, we defined a novel survival probability function, which is dependent on the time of service, emergency vehicle type, travel time, and quality of care at the receiving hospital. It is of interest to observe what policies perform better when there are multiple hospitals available in a region and what assignment behaviors increase the expected number of survivors. We study the results under a number of scenarios for each resource element.

Chapter 2

Optimal service order for mass-casualty incident response

Abstract

In the aftermath of a mass-casualty incident, one of the first steps in the response is to triage the casualties. Triage systems categorize the casualties based on criticality, and then prioritize casualties for transfer to hospitals for further treatment. The prioritization is usually based on simply ordering the casualty types without considering the available resources to transport them and the scale of the disaster. These factors can significantly affect the outcome of the rescue efforts. In this research we study a mathematical model to incorporate the above mentioned factors in the triage process. We assume a disaster location with a set of casualties, categorized by criticality and care requirements, that must be transported to hospitals in the region using a fleet of available ambulances. The goal is to maximize the expected number of survivors. We analyze the structure of the optimal solution to this problem, and compare the performance of the model with the current practice and other related models in the literature.

2.1 Introduction

In this paper we study the response to mass-casualty incidents (MCI). Unlike other emergencies, MCIs are of sufficient scale that the decision-maker (i.e., the incident commander) must manage the response under limited resources (e.g., medical transport). Thus, planning and management are critical to allocate limited resources as efficiently as possible (Antommaria et al., 2011). After an MCI, the first responders initially gather and triage the casualties. Triage is the process of categorizing casualties based on the severity of their injuries, and is an essential MCI management tool; its purpose is to use the available resources in the most efficient manner.

There are several triage methods for MCIs, including START (Simple Triage and Rapid Treatment Method), Homebush, Triage Sieve, Sacco Triage Method (STM), and CESIRA

(Lerner et al., 2008) (hospital emergency departments have their own triage methods, see Iserson and Moskop, 2007, for details). For most triage methods there is an implied service order for transporting casualties to hospitals, which is based on the triage categories and does not account for the available resources (e.g., ambulances), the scale of the incident, or the casualty mix. For instance, START, one of the most common triage method in the US (Cone and MacMillan, 2005), has four color-coded categories: immediate (red) for the most critical casualties that need attention within an hour, delayed (yellow) for serious injuries, but not expected to deteriorate for several hours, minor (green) for victims with relatively minor injuries, and expectant (black) for victims unlikely to survive (Benson et al., 1996). START has an implied service order of most critical to least critical (i.e., red, yellow, and then green), regardless of the situation. A review of the literature reveals that there has been no scientific evaluation of effectiveness of START or any of the other common triage methods (Jenkins et al., 2008; Lerner et al., 2008). One measure of triage effectiveness is the survival rate among the casualties that are treatable (e.g., those in the red, yellow, and green START categories) in the hours to days following the incident. The *research question* we are studying is *"how can we find a service order that maximizes the survival rate of an MCI?"*

This research question was motivated in part by our participation in a Full Scale Exercise (FSE) conducted by a regional airport, which simulated a plane crash with a large number of casualties (played by moulaged volunteers). The simulated response included the triage and transport of casualties to nearby hospitals. The triage method used was START. Of the nearby hospitals, one is designated as a Level 1 Trauma Center (the highest level of treatment), which is best able to effectively treat the more critical casualties, while the other hospitals are well suited for less critical casualties. Papers that study the service order problem faced in the FSE using operations research include Sacco et al. (2005); Mills et al. (2011, 2013). We discuss the models proposed in these papers in more details later, but they assume the service time (which is based on transporting the casualty to a local hospital) is the same for all causality types. This assumption fits a single hospital problem. But there

are often multiple hospitals, where certain hospitals are better suited to particular casualty types, thus yielding different service time for the casualty types.

Sacco et al. (2005); Mills et al. (2011, 2013); Dean and Nair (2014) assume casualty deaths occur after service (e.g., at the hospital), and the survival probability decreases as service is delayed. Conversely, Argon et al. (2008, 2011); Jacobson et al. (2012) study the service order problem under the assumption that casualty deaths occur before service. Jacobson et al. (2012) uses a Markovian model to study the service order problem, under the assumptions that lifetimes and service times are exponentially distributed (no previous study has shown that exponential distribution is a well-suited distribution for lifetime or service time in an emergency response effort). Based on our reading of the medical and emergency management literature, casualty deaths are more likely to occur after service (as modeled in Sacco et al., 2005; Mills et al., 2011, 2013; Dean and Nair, 2014) rather than before service (as modeled in Argon et al., 2008, 2011; Jacobson et al., 2012), at least for the relevant casualty types (i.e., not expectant in START). For instance, Frykberg and Tepas III (1988) studied 220 terrorist bombing incidents with 2,934 casualties surviving the attack and waiting for service. They recorded 40 subsequent deaths among the 2,934 survivors, of which only one occurred prior to receiving service. Sacco et al. (2005) generate survival probabilities for different casualty types (based on the STM triage system) from historical trauma data, which is used in a model that assumes casualty deaths occur after service.

Li and Glazebrook (2011) study the service order problem considering imperfect classification of casualties using a Bayesian approach in dealing with uncertainties and develop heuristics approaches for the problem. There are two types of misclassification: assignment of a more-critical casualty to a less-critical type, *undertriage*, and assignment of a less-critical casualty to a more-critical type, *overtriage*. Frykberg (2002) state that undertriage could lead to preventable deaths and it should be avoided in both daily and mass-casualty triage operations, but state no reported case of undertriage exists in the past bombing attacks. In dealing with daily trauma patients, overtriage occurs at a rate of around 50% (Kreis Jr et al., 1988), which can also be necessary to avoid preventable deaths by reducing undertriage as

much as possible (Frykberg, 2002). On the other hand, Frykberg and Tepas III (1988) conclude overtriage could be as deadly as undertriage in a mass-casualty incident. Overtriage of less-critical casualties could delay service for more-critical casualties and potentially threaten their survival. In contrast, Hupert et al. (2007) find that overtriage could have mixed effects on outcome of triage operation through simulation analysis. In this paper, we study the optimal triage service order problem under the more realistic assumption that casualties do not die before service (but they do deteriorate, thus decreasing their survival probability); triage misclassification is beyond the scope of this paper.

Currently, triage decisions are made based on fixed rules, regardless of the available resources and the number or mix of casualties. The main contributions of this paper can be summarized as follows: **1)** we develop a tractable model that determines the optimal service order considering the mix of casualties, casualty type dependent service times, multiple servers, and more than two casualty types; **2)** we compare our model with others in the literature, considering both structural properties and performance; **3)** we identify structural properties of the optimal solution, extending and generalizing the work in other papers; and, **4)** we discuss important data issues, and more realistic problems, including multiple servers and casualty types.

The remainder of this paper is organized as follows. In §2, we describe the problem and provide our model formulation. We also analyze some of the characteristics of the model. Moreover, we provide formulations for other models from the literature and analyze their assumptions and characteristics. In §3, we study a simplified case of the problem with two casualty types and multiple servers to gain insight into the structure of the optimal solution. We analyze how survival probability distribution and an arbitrary number of casualty types affect the optimal solution and the results in §4 and §5, respectively. Finally, conclusions and future works are discussed in §6.

2.2 Model Description

Consider a mass casualty incident (MCI). First responders triage the casualties and place them into one of the casualty types in the ordered set P, where type 1 is the most critical and type $|P|$ the least critical. For instance, in START type 1 would be immediate casualties (identified with red), type 2 delayed casualties (identified with yellow), and type 3 for casualties with minor injuries (green). The number of casualties of type $i \in P$ are denoted by n_i. We assume that all casualties are available for service at time zero. There are N_t servers (e.g., an ambulance and crew) available at time t (this parameter usually increases over time as more resources arrive from surrounding areas), which work in a non-preemptive manner. Service times, $s_i,\ i \in P$, include any care required to stabilize the casualty and the round trip travel time to an appropriate hospital. Because travel time represents a significant part of the service time, higher criticality does not necessarily imply longer service times. In the motivating FSE (see § 1) the hospital that is designated as a level 1 Trauma Center (which represents the highest level of trauma care), would be more appropriate for the more critical casualties, while other hospitals in the vicinity would be appropriate for less critical casualties. Thus the travel times (and thus service times) depend on the location of the MCI, in relation to the various hospitals and there abilities. The objective is to determine a service order that maximizes the number of survivors, given survival probability functions f_{it} for each casualty type i when service is delayed until time t. (See, for instance, Sacco et al., 2005; Mills et al., 2013, which provides survival curves derived from data.).

First, we introduce Model 1, an integer programming formulation for the service order problem. The definitions of the sets and parameters, decision variables, and formulation follow. To model this problem we divide the time horizon into T time intervals, and assume the data is discrete.

Sets and Parameters:

P set of casualty types in the triage system

n_i number of casualties of type i at the beginning of time interval 1, $\forall i \in P$

s_i number of time intervals it takes to serve a casualty of type i, $\forall i \in P$

T number of time intervals required to serve all the casualties, $\sum_{i \in P} n_i s_i$

f_{it} survival probability for a casualty of type i that is served in time interval t, $\forall i \in P$, $t = 1, \ldots, T$

N_t number of servers available at time interval t, $t = 1, \ldots, T$

Decision Variable:

x_{it} number of casualties of type i, whose service starts at time t, $\forall i \in P$, $t = 1, \ldots, T$

$$\textbf{Model 1:} \quad \max \quad \sum_{i \in P} \sum_{t=1}^{T} f_{it} x_{it} \tag{2.1}$$

$$\text{s.t.} \quad \sum_{t=1}^{T} x_{it} = n_i, \quad \forall i \in P \tag{2.2}$$

$$\sum_{i \in P} \sum_{f=1}^{min(s_i,t)} x_{i(t-f+1)} \leq N_t, \quad t = 1, \ldots, T \tag{2.3}$$

$$x_{it} \geq 0 \text{ and integer}, \ \forall i \in P, \ t = 1, \ldots, T \tag{2.4}$$

Objective function (2.1) maximizes the sum of the survival probabilities of all casualties at the time of service. We assume emergency personnel will perform required interventions to stabilize the casualty when the service begins. Thus, the decrease is survival probability rate during the service period is not considered. Constraint (2.2) guarantees that all the casualties are served. Constraint (2.3) limits the number of active servers to the number of servers available. Constraint (2.4) is the integer and non-negativity constraint.

Proposition 1 *For Model 1 with a continuous relaxation of the x-variables, there exists an optimal solution in which the x-variables have integer values.*

Proof. The coefficient matrix A, has $|P| + T$ rows and $T|P|$ columns. We show that A is totally unimodular and use this to prove the proposition. Order the columns (variables) of A by casualty-type from 1 to $|P|$, and within each type, by time intervals from 1 to T. The rows are ordered Constraint (2.2) from 1 to $|P|$, and then Constraint (2.3) from 1 to T. This yields an A matrix like the one in Figure 2.1, where in sub-matrix $i \in P$, all T elements in row i are 1 and rest of the rows are all 0, while for rows $i \in \{|P| + 1, \ldots, |P| + T\}$ all elements are zero, except for $\max\{1, i - |P| - s_i + 1\}, \ldots, i - |P|$, which are 1 (in the figure, non-zero coefficients are displayed, everything else is zero, and we set $s_1 = 2$, $s_2 = 3$, and $s_{|P|} = 4$, arbitrarily for illustration purposes). Given this form of A, we perform the following

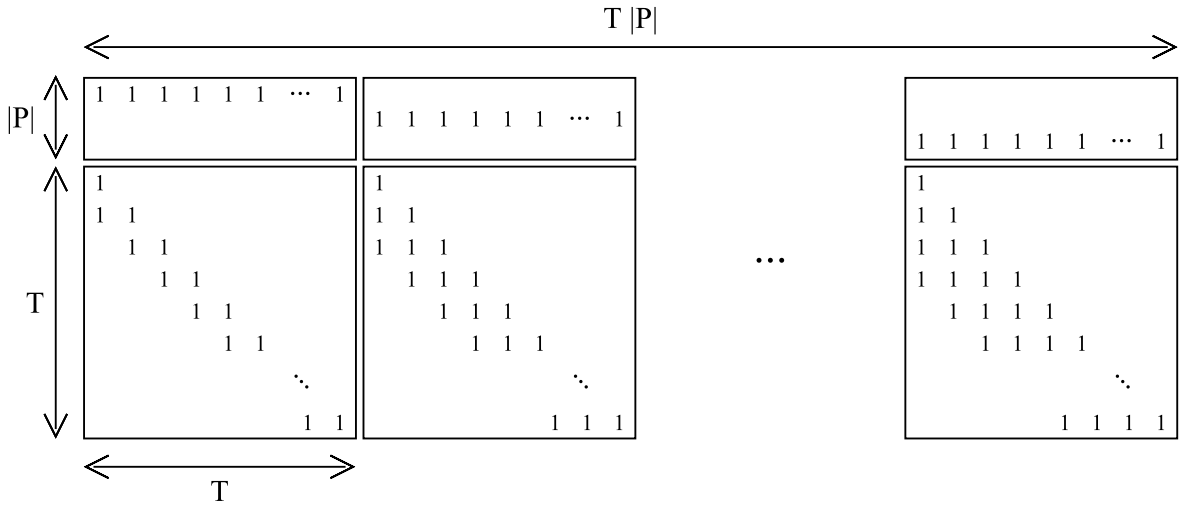


Figure 2.1: Structure of the coefficients matrix, A, for Model 1

basic matrix operations. For each sub-matrix $i \in P$ we subtract column $j = T, \cdots, 2$ from columns $k = j - 1, \cdots, \max\{0, j - s_i + 1\}$. Next we multiply the first $|P|$ rows by -1. The ordering and column operations used do not affect total unimodularity (Schrijver, 1998), and now the A matrix is such that each column has at most two non-zero values, one $+1$

and one -1, which indicates that A is totally unimodular. Given that the right-hand-side values are integer, by Cramer's Rule, all x decision variables will be integer-valued at each extreme point (Bazaraa et al., 2011), and the LP relaxation will have an integer valued optimal solution. $\square$

Next, we present other models from the literature using our notation (for easier comparison), these models all represent special cases of Model 1 for more narrowly defined problems. Sacco et al. (2005, 2007) present a linear program for the triage service order problem:

$$\textbf{Model 2:} \quad \max \quad \sum_{i \in P} \sum_{t=1}^{T} f_{it} x_{it} \tag{2.5}$$

$$\text{s.t.} \quad \sum_{t=1}^{T} x_{it} = n_i, \quad \forall i \in P \tag{2.6}$$

$$\sum_{i \in P} x_{it} \leq N_t, \quad t = 1, \ldots, T \tag{2.7}$$

$$x_{it} \geq 0, \quad \forall i \in P, \; t = 1, \ldots, T. \tag{2.8}$$

Objective function (2.5) is identical to (2.1), while Constraint (2.6) ensures that all the casualties are served. Constraint (2.7) limits the number of servers to those available, and Constraint (2.8) is the integer and the logical non-negativity constraint. This LP is similar to Model 1, except that it implicitly assumes service times are equal for each casualty type, which is a limitation. By Proposition 1, this LP has integer optimal solutions, much like Model 1.

Next, is Model 3A, proposed by Mills et al. (2013), to study the service order problem with one server. Model 3A also assumes that service times for all casualty types are equal, and for simplicity sets them all to one time interval (e.g., service time is the transportation time from the disaster location to the receiving hospital), thus $T = \sum_{i \in P} n_i$.

$$\textbf{Model 3A:} \quad \max_{W} \quad \sum_{i \in P} \int_{W(i)} f_i(t)dt \tag{2.9}$$

$$\text{s.t.} \quad \mu(W(i)) = n_i, \quad i \in P \tag{2.10}$$

$$\bigcup_{i \in P} W(i) = [0, T], \tag{2.11}$$

$$W(i) \cap W(j) = \emptyset, \quad \forall i, j \in P, i \neq j, \tag{2.12}$$

where W is a set-valued decision variable such that $W = \{W(i), i \in P\}$, and $W(i)$ identifies continuous time intervals in which service is allocated to casualty type $i \in P$. The objective function (2.9) maximizes the sum of survival probabilities, $f_i(t), i \in P$, for all casualty types. Constraint (2.10) guarantees that total number of the casualties served over the time horizon, $\mu(W(i)),\ i \in P$, is equal to total number of casualties in each category, $n_i, i \in P$. Constraints (2.11) and (2.12) ensure that service times for all casualty types do not exceed the time horizon T or overlap, respectively. Model 3A, while not a linear program, is equivalent to Model 2 restricted to a single server. Mills et al. (2013) show that there exists an optimal solution to Model 3A where only one casualty type is served at any given time. This result follows from Proposition 1, which shows that there exists an integral optimal solution for the continuous relaxation of the problem, and restricting the problem to a single server. While Model 3A allows for an arbitrary number of casualty types, Mills et al. (2013) simplifies Model 3A for a special case having only two casualty types (thus $T = n_1 + n_2$) to gain insight into the structure of the optimal solution. Furthermore, the use of two casualty types is of practical interest as it represents the two casualty types from START that are most significant to the problem (i.e., immediate and delayed). This simplification, denoted as Model 3B, uses an alternate objective function based on the function $g(t) = f_2(t) - f_1(t)$.

$$\textbf{Model 3B:} \quad \max \quad \int_{W(2)} g(t)dt + C \tag{2.13}$$

$$\text{s.t.} \quad W(2) = n_2, \tag{2.14}$$

$$W(2) \subseteq [0, T]. \tag{2.15}$$

Objective function (2.13) maximizes the survival probability for two casualty types, here C is a constant defined as $C \equiv \int_0^T f_1(t)dt$, which is equivalent to (2.9) for two casualty types. The constant is for the survival probability of type 1 casualties, to which we add the difference function $g(t)$. Constraint (2.14) ensures all type 2 casualties are served. Since the time horizon is limited to exactly the required service time for all casualties ($T = n_1 + n_2$), after the assignment of type 2 casualties, the remaining time intervals are assigned to type 1 casualties. Constraint (2.15) limits type 2 casualties service to the defined time horizon. In this model, each service requires one time interval, independent of casualty type, therefore we can easily discretize the difference function $g(t)$ for each time interval, that is, $g_t = f_{2t} - f_{1t}$, which we sum for every interval a casualty of type 2 is served, which is indicated by the binary variable x_t ($x_t = 1$ if a casualty of type 2 is served in interval t, else $x_t = 0$ and a type 1 casualty is served), which yields Model 3C (once again $T = n_1 + n_2$).

$$\textbf{Model 3C:} \quad \max \quad \sum_{t=1}^{T} g_t x_t \tag{2.16}$$

$$\text{s.t.} \quad \sum_{t=1}^{T} x_t = n_2 \tag{2.17}$$

$$x_t \in \{0, 1\}, \forall t \in 1, \cdots, T. \tag{2.18}$$

Objective function (2.16) maximizes the survival probability, and is equivalent to objective function (2.13) in Model 3B. Constraint (2.17) ensures all type 2 casualties are served, and it is equivalent to Constraint (2.14). Finally, Constraint (2.18) limits the service to the time

horizon, which is equivalent to Constraint (2.15).

Models 1 through 3C have a hierarchical relationship; Model 1 is the most general, while Models 3B and C are for the most narrowly defined problem. Model 1 can address multiple casualty types, a time varying number of servers, and a unique service time for each casualty type. Unlike Model 1, Model 2 requires equal service times for all casualty types, while Model 3A is equivalent to Model 2, but limited to a single server. Model 3B is a simplification of Model 3A with two casualty types, and Models 3B and 3C are equivalent. Table 2.1 summarizes the differences among all the models.

Table 2.1: Comparison of service order models

Feature	Model 1	Model 2	Model 3A	Model 3B	Model 3C
Casualty types	Unlimited	Unlimited	Unlimited	Two	Two
Servers	Multiple	Multiple	Single	Single	Single
Service times	Variable	Equal	Equal	Equal	Equal

In addition, Dean and Nair (2014) develop an integer programming model, which addresses some of the shortcomings of the Model 2, including accounting for hospital resources. They have parameters for initial number of servers and hospital beds and after assignment to the casualties, they become occupied for a certain amount of time and then, they become available again. Due to the added complexity, it is more difficult to develop any insight into the structure of the optimal solution and the authors suffice to some high level numerical analysis and comparison of their results with that of Sacco et al. (2005). Except hospital assignment, which is beyond the scope of this paper (we focus on prioritization of service to casualties), Model 1 is an extension of both Model 2 and the one developed by Dean and Nair (2014), as not only it considers resource reusability, it also allows for adding additional resources in later time intervals.

Addressing multiple casualty types is important, as there are often casualties that do not fit neatly into the two START classifications discussed above. For instance, pediatric casualties benefit from specialized care and services that should be considered in the management of an MCI. In fact, pediatric casualties have their own triage system, as well as hospitals

that specialize in children, and children often represent a significant number of the MCI casualties (Lyle et al., 2009; Mace and Bern, 2007). Because each casualty type has different service and care requirements, the expected service times vary among different casualty types and models should be able to address this difference in order to generate practical results. Another issue ignored by the models in the literature is how number of servers could vary through time during response efforts. It is likely that there are a few servers at the disaster location initially, but their numbers could potentially be augmented through time as part of the regional response. Model 1 captures all these essential elements of an MCI response. Despite this, we first look at the simplified problem with two casualty types, to study the structural solution properties of this problem.

2.3 Survival Data

In this section we discuss the survival probability functions used in the numerical analysis, which are based on the scaled log-logistic survival functions derived in Mills et al. (2013) from data in Sacco et al. (2005) and discuss data issues. The survival probability function is as follows:

$$f_{it} = \frac{\beta_{0,i}}{(t/\beta_{1,i})^{\beta_{2,i}} + 1}, \quad \forall\, i \in \{1, 2\}, \tag{2.19}$$

where $\beta_{0,i}$, scales the function to the initial survival probability for casualty type i, and $\beta_{1,i}$ and $\beta_{2,i}$ are parameters of the log-logistic distribution. Sacco et al. (2005) derive survival probabilities for a 13 category triage system using a retrospective analysis of data from 76,459 casualties from trauma centers in Pennsylvania, where category 0 has the lowest survival probability and 12 the highest. Then, using a Delphi technique, the transition from less critical to more critical categories through time were estimated. Mills et al. (2013) produce five scenarios where the 13 casualty categories from Sacco et al. (2005) are probabilistically mapped into the two START categories of interest (see Mills et al., 2013, for details); while Sacco et al. (2005) assume a transition through the casualty types due to delay in service,

Mills et al. (2013) have a survival probability function for each casualty type, and no transitioning of casualties between types. Scenario 1 is the most time-sensitive (lower survival probabilities) and Scenario 5 is the least time-sensitive.

Table 2.2: Parameters for five survival probability scenarios from Mills et al. (2013) for survival probability function (2.19) for 5-minute time intervals

	Type 1			Type 2		
Scenario	$\beta_{0,1}$	$\beta_{1,1}$	$\beta_{2,1}$	$\beta_{0,2}$	$\beta_{1,2}$	$\beta_{2,2}$
1	0.09	17	1.01	0.57	61	2.03
2	0.15	28	1.38	0.65	86	2.11
3	0.24	47	1.30	0.76	138	2.17
4	0.40	59	1.47	0.77	140	2.29
5	0.56	91	1.58	0.81	160	2.41

Much of the analysis and heuristic development in Mills et al. (2013) is based on the following assumption on the relationship between the two survival probability functions, f_{it}, $i \in 1, 2$.

Assumption 1 *There exists a time t_m such that $f'_{1t} < f'_{2t}$ for $t < t_m$, $f'_{1t_m} = f'_{2t_m}$, and $f'_{1t} > f'_{2t}$ for $t > t_m$.*

Assumption 1 indicates that, before time t_m, the survival probability of type 1 casualties, f_{1t}, initially decreases at a faster pace than the survival probability of type 2 casualties, but after time t_m, the type 2 survival probability starts to decrease faster than the survival probability of type 1. Mills et al. (2013) state that the survival probability functions defined in (2.19) using the data from Table 2.2 adhere to Assumption 1. This data is for 5-minute time intervals, and throughout this paper we transform survival probability function (2.19) to account for this.

These five scenarios are problematic in two ways. First, consider these scenarios from the triage perspective. Triage methods place casualties into types based on a (simple) algorithm. For example, based on the START algorithm, where classification is based on pulse rate, respiratory rate, and responsiveness of victims (Benson et al., 1996), there is no way for the

first responder to know which survival probability curves are applicable. Under Scenario 1 type 2 casualties have an initial survival probability of 0.57 and an expected lifetime (or more precisely, the time before which service is no longer useful) of 269.1 minutes, while under Scenario 5 type 1 casualties have an initial survival probability of 0.56 and an expected lifetime of 554.3 minutes. Thus, the Scenario 1 type 2 casualties, while having a slightly higher initial survival probability than the Scenario 5 type 1 casualties, are the more time critical, i.e., their survival probability drops faster (thus the lower expected lifetime). Also, the initial survival probabilities for type 1 casualties range from 0.09 to 0.56. Thus, we observe a wide variance between casualties of the same type in different scenarios, and a large overlap in criticality between the two casualty types in the different scenarios. Given that the survival probability curves are crucial in determining the optimal response, any modeling effort would require more stable survival probability curves for each casualty type. These scenarios were produced by probabilistically mapping the 13 casualty types (based on a more complex algorithm) from Sacco et al. (2005) into two casualty types (see Mills et al., 2013, for details). This does not necessarily reflect the START algorithm, but the stability of any triage algorithm is beyond the scope of this paper.

The other issue with the five scenarios is the shape of the survival probability curves, specifically as defined by the initial survival probability. For example, consider Scenario 2. For a type 1 casualties the initial survival probability is $\beta_{0,1} = 0.15$, which results is a flattened log-logistic survival probability curve having an expected lifetime of 62.8 minutes, that represents casualties that having a low survival probability, but are not very *time-critical* (the low expected lifetime somewhat reflects low initial survival probability, i.e., 85% of casualties will eventually die no matter how fast service is provided, and the remaining 15% have a longer life expectancy). The type 2 casualties have a higher probability of survival (initially 0.65), and an expected lifetime of 417.6 minutes. This increases the relative time criticality of the type 2 casualties as their survival probability has much more opportunity to drop.

Relating these two issues and again considering START, should casualties that have, at best, a 9% or 15% chance of survival be classified as immediate (i.e., red, and given the highest

priority) or expectant (i.e., black, and given the lowest)? Thus, we have identified two areas for further study, first is the quality of the triage algorithm, for instance, if the START algorithm does produce varying scenarios like those described in Scenarios 1-5, then a more accurate algorithm is required, one that better categorizes casualties. We also question the scaling values in Scenarios 1-5 (i.e., the initial survival probability), which, if too low, makes giving type 1 casualties higher priority less desirable. We do not argue that the proposed survival probability data and function proposed by Mills et al. (2013) and Sacco et al. (2005) are either a good or a bad fit, and in this paper we use the five scenarios for comparison purposes.

2.4 Two Casualty Types, Multiple Servers and Unequal Service Times

In this section, we study the structure of the optimal solution to the triage service order problem with two casualty types and multiple servers. Considering only two casualty types helps us show analytical results, and is somewhat practical as in most common triage methods, there are typically two critical casualty types competing for resources. Again, for START these two types are immediate and delayed. The minor category can wait longer before receiving service, while the expectant category has a very low survival probability, regardless of availability of the resources. We extend the work in Mills et al. (2013) by considering multiple servers and more importantly, potentially unequal service times for the two casualty types, and provide alternative solution algorithms to this more general problem. In our analysis, we consider two simple strategies, a strategy where all type 1 casualties are served first, followed by the type 2 casualties, which we denote as $S(1,2)$; this is the implied service order of the START triage system, and we also consider the opposite order, which we denote as $S(2,1)$.

Assumption 1 uses derivatives to identify t_m. Instead it is more appropriate to use the

change in survival probability over the service time (which is more accurate, and important for unequal service times). Given this, we define the following parameter for the change in survival probability:

Definition 1 $f_{it}^{\delta} \equiv f_{it} - f_{it+\delta}$.

f_{it}^{δ} denotes the decrease in survival probability starting at time t over the next $\delta > 0$ time intervals. Given this definition, we make the following assumption:

Assumption 2 *There exists a time* t_s *such that* $f_{1t}^{s_2} > f_{2t}^{s_1}$ *for all* $t < t_s$, $f_{1t_s}^{s_2} = f_{2t_s}^{s_1}$, *and* $f_{1t}^{s_2} < f_{2t}^{s_1}$ *for all* $t > t_s$.

Assumption 2 is a generalization of Assumption 1 that allows for unequal service times for two casualty types, and survival functions that adheres to the latter, also adhere to Assumption 2, and vise versa. When a type 1 casualty is being served, service to a type 2 casualty is delayed by s_1 time intervals (i.e., until service to type 1 is finished). Thus, decrease in type 2 survival probability is $f_{2t}^{s_1}$, and using the same logic, decrease in survival probability for type 1 is $f_{1t}^{s_2}$.

When service times are equal ($s_1 = s_2$), Mills et al. (2013) shows that for one server ($m = 1$) and survival probabilities that adhere to Assumption 2, there exists an optimal solution structured such that *service of type 2 casualties form a continuous interval* which allows optimal solutions to be described using three time intervals, the first interval for serving n_1', $0 \leq n_1' \leq n_1$ type 1 casualties, the second for serving all n_2 casualties, and third intervals for serving the remaining $n_1 - n_1'$ type 1 casualties. Using this optimal structure, Mills et al. (2013) proposes two heuristics for Model 3B (equal service times and one server). One heuristic, denoted QS-ReSTART, simply decides on either $S(1,2)$ or $S(2,1)$ as the service order plan, while the second heuristic, denoted QD-ReSTART, provides a plan that is either $S(1,2)$, $S(2,1)$, or time-dependent, which serves type 1 casualties until the estimated

switching point, τ, after which service switches from type 1 to type 2. If $\tau \leq 0$, QD-ReSTART results in $S(2,1)$, if $\tau \geq n_1 s_1$ it results in $S(1,2)$.

Next, we introduce Algorithm 1 that produces optimal solutions when service times are equal ($s_1 = s_2$), even considering m servers and survival probabilities that **do not necessarily adhere** to Assumption 2. To describe this greedy knapsack algorithm, we assume, without loss of generality, that service times are one unit. Start with a list ORDER of n_1+n_2 items, where each item has four elements: decision epoch (service time, t), server, $g_t = f_{2t} - f_{1t}$, and casualty type. ORDER$[i]_j$ refers to the element j of item i in ORDER, for all $i \in \{1, \ldots, n_1+n_2\}$ and $j \in \{1,2,3,4\}$. The first *for-loop* sets all elements of ORDER with the decision epochs, the servers, the g_t value for the epoch, and sets the casualty type to 1 (for type 1). Next, ORDER is sorted on g_t (third element) from largest to smallest. Finally, the first n_2 items, those having the highest g_t-values, are re-set to type 2 casualties, that is, the fourth element in ORDER is set to 2. ORDER now indicates the optimal service order and server for each casualty.

Algorithm 1: Optimal algorithm for $|P| = 2$, $s_1 = s_2$, and m servers

```
1  begin
2      Create empty indexed-list ORDER;
3      SERVER ← 1;
4      for i ← 1 to n_1 + n_2 do
5          t ← ⌊(i−1)/m⌋;
6          ORDER[i] ← (t, SERVER, g_t, 1);
7          SERVER ← SERVER + 1;
8          if SERVER > m then
9              SERVER ← 1
10     sort ORDER from largest to smallest (descending) value of the third element (g_t);
11     for i ← 1 to n_1 + n_2 do
12         if i ≤ n_2 then
13             ORDER[i]_4 ← 2
```

Of course, Algorithm 1 provides the optimal solution to Models 3B and 3C when $m = 1$, and is not dependent on Assumption 2, unlike QD-ReSTART and QS-ReSTART, and can be very efficiently solved. Still, the optimal structure under Assumption 2 (described above) is clear in the context of this algorithm. Here each of the $n_1 + n_2$ "items" has the same weight ($s_1 = s_2 = 1$) and values based on g_t. By Assumption 2 g_t has one maximum, and choosing the n_2 time intervals with the largest g_t values to serve type 2 casualties is the optimal solution, and they form a continuous interval. When Assumption 2 does not hold, service to type 2 casualties does not necessarily form a continuous interval.

Next, we consider *unequal service times* and *multiple servers.* In Mills et al. (2013) the rational for equal service times is that they are based on the time required to transport a casualty to the hospital. But often there are multiple hospitals to consider, and sending all the casualties to the closest hospital is often not be a good idea, instead this decision should be based on the capabilities of each of the regional hospitals (e.g., a Level 1 Trauma center is best able to handle more critical patients). Thus, the location of the MCI and the corresponding hospitals would determine the service times. Also, the time required for stabilization of different casualty type could vary, which also leads to unequal service times. We now offer a generalization of the structural results from Mills et al. (2013).

Proposition 2 *There exists an optimal solution to Model 1 under Assumption 2, where service of type 2 casualties once started, is not interrupted in any of the servers.*

Proof. See Appendix.

Proposition 2 shows that with unequal service times, the optimal solution for each server has a similar structure to that with equal service times, and that service is balanced across the servers. We serve type 2 casualties in an interval containing t_s, assuming there are sufficient casualties. Increasing the number of servers does not change t_s, but it does change t_i^* (the time for server i to switch from type 1 to type 2 casualties). As the number of

servers increases, t_i^* either remains the same, or increases. Having multiple servers favors the $S(1,2)$ solution. Consider the interval in which type 2 casualties are being served (depending on t_s) when there is one server. Adding another server reduces this interval (still around t_s, assuming sufficient number of casualties), potentially increasing the length of the first interval serving type 1 casualties, and of course more type 1 casualties will be served in this first interval because of the increased number of servers, which we see in the results displayed in Table 2.3, from Model 1, $S(1,2)$, and $S(2,1)$ strategies for two casualty types and different number of servers. Base on the same logic, as we add more servers, more type 1 casualties are served initially (assuming there is a sufficient number of them), as t_s remains the same and type 2 service gets more condensed around that point. If there are insufficient number of type 1 casualties, we finish service to them and start serving type 2 casualties earlier. In this case, the optimal service order becomes $S(1,2)$. This confirms favoring $S(1,2)$ as we increase number of the servers.

Unlike the special case where service times are equal, this is not a simple knapsack problem. There is not a well-defined set of decision epochs, rather the time of decision epochs depends on previous decisions, and thus survival probabilities in the next decision epoch depend on the casualty type that is currently being served. Despite this, if the survival probability function adheres to Assumption 2, we know the optimal structure, which allows us to use a simple search, starting with the $S(2,1)$ strategy, evaluating the expected number of survivors, and then moving one type 1 casualty to the end of the service queue on each server and again evaluate the expected number of survivors, and repeat until the expected number of survivors decreases. Algorithm 2 shows the detailed steps to find the optimal solution having two casualty types and m servers, with no restriction on service times. Building on the insight from the Proposition 2, Algorithm 2 generates the optimal service order by finding the point in which service should be switched from type 1 to type 2. The algorithm starts with $S(2,1)$ strategy on each server, while number of the casualties are balanced on each server (equal or almost equal number of type 1 and type 2 casualties on each server). The solution is then presented in terms of three service intervals on server j, $n'_{1,j}$ type 1, $n_{2,j}$

type 2, and $n_{1,j} - n'_{1,j}$ type 1. If $n'_{1,j} = n_1$ for all $j \in \{1, \ldots, m\}$, solution is the same as $S(1,2)$ and if $n'_{1,j} = 0$ for all $j \in \{1, \ldots, m\}$, the solution is $S(2,1)$.

Algorithm 2: Optimal algorithm for $|P| = 2$, $s_1 \neq s_2$, and m servers

```
1  begin
2      n'_{1,j} ← 0 for all j ∈ {1, ..., m};
3      SERVER ← 1;
4      TS_NEW ← total expected survivors for the initial strategy, S(2,1);
5      TS_OLD ← 0;
6      while strategy is not S(1,2) AND TS_OLD < TS_NEW do
7          move one type 1 casualty to the beginning of queue for server SERVER;
8          SERVER ← SERVER + 1;
9          if SERVER > m then
10             SERVER ← 1
11         TS_OLD ← TS_NEW;
12         TS_NEW ← total expected survivors for the new strategy;
13         if TS_OLD < TS_NEW then
14             n'_{1,SERVER} ← n'_{1,SERVER} + 1;
```

Result: optimal solution is to serve $n'_{1,j}$ type 1, $n_{2,j}$ type 2, and $n_{1,j} - n'_{1,j}$ type 1 casualties on server j for all $j \in \{1, \ldots, m\}$

Figure 2.2 displays f_{it}^s for Scenario 5 (see Table 2.2) where $s = s1 = s2 = 30$ minutes. The dashed curve (red) is the change in type 1 survival probability function over the 30 minute service time and the solid curve (yellow) is for type 2 casualties. As expected, both functions are decreasing with type 1 initially decreasing at a faster pace until point $t_s = 341$ (see Assumption 1), after which the type 2 function decreases faster. Given sufficient number of casualties, the interval in which type 2 casualties are served contains t_s in the optimal

solution. For example, in Figure 2.2, the optimal solution for six cases, which vary in the mix of casualties, is illustrated. For each of these cases, a bar represents the service decisions, the hatched section (red) indicates service to type 1 casualties and solid (yellow) section indicates service to type 2. In each case, there are total of 50 casualties and number of the casualties served in each interval is stated on the graph. The graph illustrates that as we increase number of type 2 casualties, their service interval grows around t_s and their service is not interrupted.

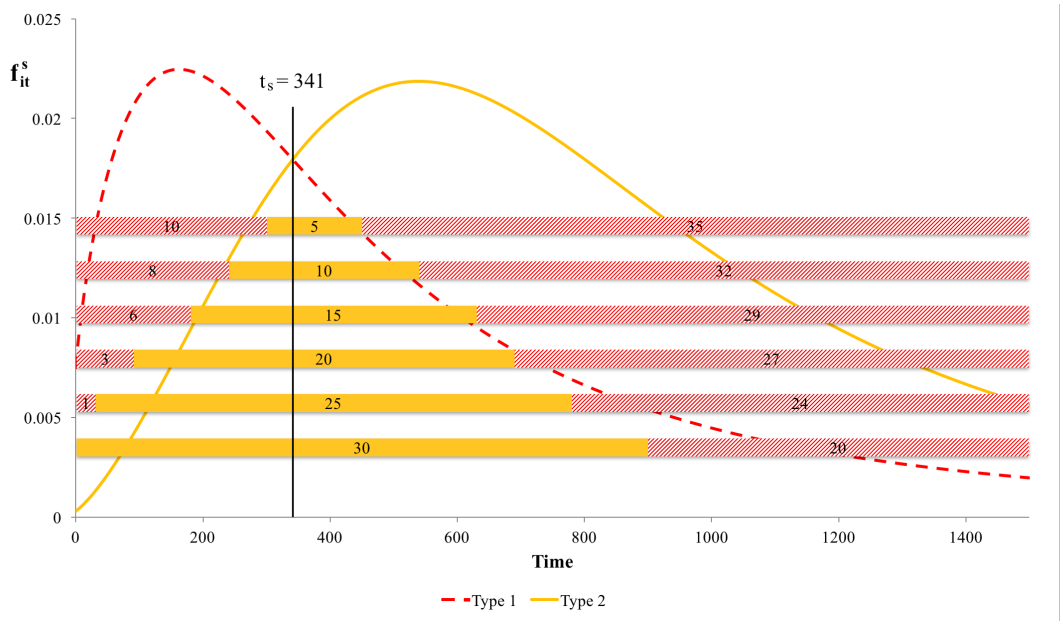


Figure 2.2: Optimal service order for six different mixes of casualties with $s = s1 = s2 = 30$ for Scenario 5

Figure 2.3 illustrates optimal solutions from Model 1 when service times are not equal ($(s_1, s_2) = (35, 25)$, both in minutes) under Scenario 5, these solutions adhere to Proposition 2. The dashed curve (red) is $f_{1t}^{s_2}$, the change in survival probability for type 1 casualties per service time of casualty type 2, and solid curve (yellow) is $f_{2t}^{s_1}$ for casualty type 2. As in Figure 2.2, hatched intervals (red) indicate service of type 1 casualties, and the solid intervals (yellow) indicate service of type 2 casualties. Comparing Figures 2.2 and 2.3, we can see increasing type 1 service and decreasing type 2 service moves the results towards $S(2, 1)$, as

it becomes less efficient to serve type 1 casualties with greater service time and lower survival probability using limited resources compared to the type 2 casualties with shorter service time and higher survival probability.

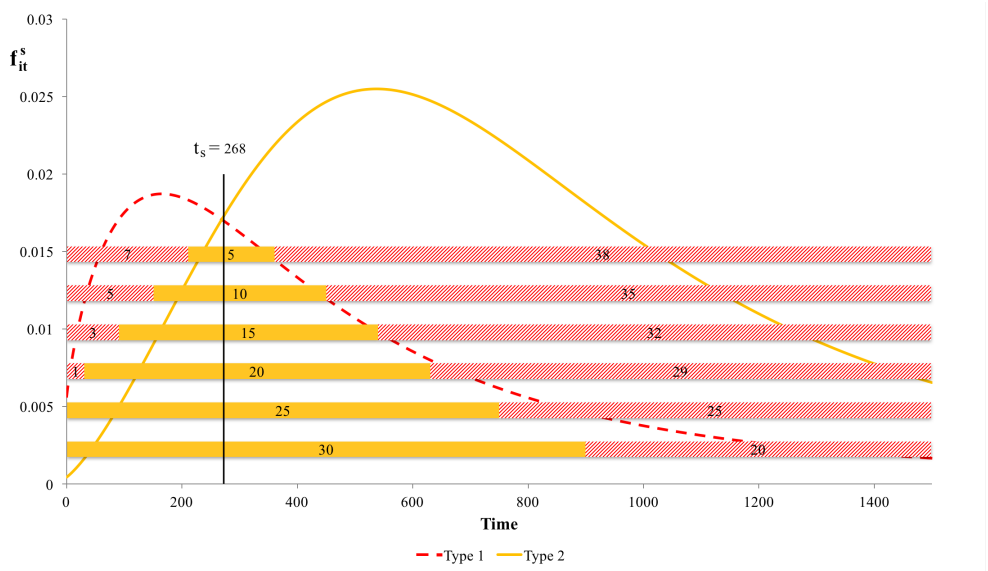


Figure 2.3: Optimal service order for different mix of casualties with $s_1 = 35$ and $s_2 = 25$ for Scenario 5

Earlier in Assumption 2 we showed that t_s is dependent on the service times. Thus, we cannot estimate the results from multiple servers by proportionally reducing service times. Based on the proposition, we can do so by proportionally reducing the number of the casualties (This does not provide the optimal solution, but provides a more accurate estimation of the optimal solution). Mills et al. (2013) approximate the solution with multiple servers by dividing service times by number of the servers, but this can cause errors as the number of servers increase. Consider Scenario 5 with $n_1 = 25$ and $n_2 = 25$ and 5 servers, each having a service time of 30 minutes. The optimal solution serves five type 1 casualties, then 25 type 2 casualties, and then the remaining 20 type 1 casualties. For this solution the expected number of survivors is 4.845 type 1, 15.804 type 2, for a total of 20.649. The approximation is obtained using one server with a service time of 6 minutes. This has an optimal solution that serves two type 1 casualties, then 25 type 2, followed by the remaining

23 type 2 casualties. For this solution the expected number of survivors is 3.454 type 1, 16.198 type 2, for a total of 19.652 survivors. This approximate method has an error of 0.997 in the number of expected survivors in this example. In addition, as stated in Assumption 2, reducing service times change t_s, which results in a different switching time t^* that changes the optimal solution (based on Proposition 2).

Table 2.3 displays the results for cases with one, two, and three servers from Model 1, $S(1,2)$. and $S(2,1)$ strategies using the survival probabilities generated from (2.19). The mean service time for each casualty type, for each simulation run, is randomly selected from a uniform distribution from 20 to 40 minutes, thus in most cases service times are unequal. Mills et al. (2013) use lognormal distribution to generate random service times for each casualty type with standard deviation of 40% of the mean, as justified from an empirical study by Ingolfsson et al. (2008) as a good fit for ambulance travel times. Thus, the randomly selected means are used to generate lognormally distributed service times for each casualty in each simulation run. All instances have 50 casualties, the number of type 1 casualties is uniformly distributed between 10 and 40, and the remaining casualties are type 2. Each scenario is simulated $5,000$ times and we report the number of times each method generates the policy having the highest objective function value (reported as *Times Best* in the table), as well as the the percent of the casualties of type 1, type 2, and total, that survive.

We observe from the results in Table 2.3 that increasing the number of servers leads to an increase in total survivors and moves the optimal solution closer to S(1,2), as shown analytically earlier. In more critical scenarios (e.g., Scenario 1) with fewer servers $S(1,2)$ is very unlikely to be the optimal service order policy, but shorter randomly generated type 2 lifetimes or type 1 service times, longer randomly generated type 1 lifetimes or type 2 service times, or a combination of all could result in $S(1,2)$ being the optimal solution. If the optimal solution is $S(1,2)$, increasing servers does not change this. On the other hand, if a solution is $S(2,1)$, decreasing servers does does not affect the order.

Next, in Table 2.4, we perform a similar numerical analysis with more realistic initial survival

Table 2.3: Simulation of the results from Model 1, $S(1,2)$, and $S(2,1)$ with different number of servers and 5-minute time intervals

	1 Server				2 Servers				3 Servers			
		Percentage Survived				Percentage Survived				Percentage Survived		
Method	Best	Type 1	Type 2	Total	Best	Type 1	Type 2	Total	Best	Type 1	Type 2	Total
						Scenario 1						
Model 1	4,794	0.68	28.97	13.68	4,635	1.33	41.73	20.50	4,557	1.82	47.43	23.88
S(1,2)	110	2.36	5.16	3.92	217	3.60	16.07	10.19	316	4.28	24.54	14.86
S(2,1)	4,773	0.68	28.95	13.67	4,630	1.33	41.75	20.49	4,556	1.76	47.47	23.87
						Scenario 2						
Model 1	4,760	0.95	40.55	19.41	4,378	2.92	53.12	27.25	4,165	4.06	57.80	30.39
S(1,2)	115	4.84	9.81	7.56	460	7.62	27.10	17.78	740	9.01	37.63	23.64
S(2,1)	4,759	0.94	40.56	19.40	4,339	2.41	53.69	27.20	4,088	3.51	58.55	30.31
						Scenario 3						
Model 1	4,568	3.41	58.15	29.47	3,951	8.85	67.19	37.87	3,672	12.67	70.40	41.24
S(1,2)	206	10.82	22.55	17.06	1,052	15.27	47.05	31.54	1,780	17.14	59.37	38.26
S(2,1)	4,561	3.24	58.52	29.47	3,655	6.59	69.11	37.50	2,963	9.06	72.82	40.52
						Scenario 4						
Model 1	4,095	7.61	57.65	31.74	3,567	19.85	66.69	43.07	3,205	27.05	69.15	47.68
S(1,2)	455	20.05	22.82	21.59	1,529	27.72	48.77	38.43	2,374	31.49	60.94	46.12
S(2,1)	3,930	5.65	59.96	31.54	2,608	12.53	71.21	41.61	1,999	17.76	74.18	45.76
						Scenario 5						
Model 1	3,640	19.22	60.18	39.09	3,330	37.60	69.66	53.39	3,294	46.11	72.89	59.23
S(1,2)	883	34.70	28.24	31.45	2,183	44.82	56.51	50.70	3,117	49.07	68.40	58.69
S(2,1)	3,189	12.42	66.24	38.36	2,042	25.29	76.48	50.92	1,527	33.18	78.97	56.28

probabilities (as discussed in § 2). In all the scenarios, we use $(\beta_{0,1}, \beta_{0,2}) = (0.8, 0.95)$ with the rest of the parameters the same as in Table 2.2. The simulation setting is the same as before.

Table 2.4: Simulation of the results from Model 1, $S(1,2)$, and $S(2,1)$, with different number of servers, revised data, and 5-minute time intervals

	1 Server				2 Servers				3 Servers			
		Percentage Survived				Percentage Survived				Percentage Survived		
Method	Best	Type 1	Type 2	Total	Best	Type 1	Type 2	Total	Best	Type 1	Type 2	Total
						Scenario 1						
Model 1	4,169	8.07	45.52	25.08	3,669	19.36	63.90	40.54	3122	27.99	67.44	46.72
S(1,2)	242	20.60	7.83	13.86	742	32.29	27.04	29.46	1302	37.10	43.35	40.06
S(2,1)	4,108	5.84	47.24	24.65	2,989	11.91	69.38	39.32	2299	18.30	72.95	44.65
						Scenario 2						
Model 1	3,874	8.48	53.99	29.83	3,502	27.94	68.92	47.30	3,041	37.66	70.92	53.52
S(1,2)	276	25.09	13.46	18.83	1,145	40.34	38.82	39.24	1,805	45.36	56.35	50.52
S(2,1)	3,505	4.82	58.14	29.64	2,204	12.67	77.99	44.25	1,571	21.93	78.35	49.68
						Scenario 3						
Model 1	3,600	17.24	66.42	40.80	3,460	41.87	77.52	58.77	2,969	49.18	78.51	62.92
S(1,2)	444	35.21	27.05	30.97	1,826	50.32	59.06	54.47	2,509	53.57	72.69	62.76
S(2,1)	2,944	10.48	73.24	40.22	1,520	22.44	86.90	54.25	1,211	32.46	84.43	58.18
						Scenario 4						
Model 1	3,580	19.87	65.35	41.93	3,446	47.36	76.91	61.26	2,984	54.05	79.45	65.92
S(1,2)	572	39.83	26.85	33.01	2,014	55.85	59.92	57.62	2,634	57.96	74.02	65.77
S(2,1)	2,788	10.95	73.38	41.04	1,413	25.55	87.42	56.26	1,162	37.05	84.71	60.83
						Scenario 5						
Model 1	3376	28.33	67.68	47.60	3,373	56.38	79.45	67.29	2,899	61.73	82.43	71.45
S(1,2)	923	49.79	32.16	40.63	2,388	63.89	65.70	64.68	2,894	64.95	78.30	71.39
S(2,1)	2,649	16.88	77.62	46.53	1,490	36.02	89.30	62.87	1,357	48.18	86.76	67.76

Table 2.4 demonstrates how increasing initial survival probability increases total expected survivors. As expected, increasing initial survival probability moves the results towards $S(1,2)$. This is mainly due to increase in the value of the t_s as explained in Survival Data section (e.g., the case explained in Survival Data). As we increase number of the servers, number of the type 2 casualties served around t_s in a single server decreases (since the service spreads across a greater number of servers). Thus, potentially the available time to initially serve type 1 casualties increase and results shift towards $S(1,2)$. In general, the pattern of change in the results remain similar to that Table 2.3. In some cases, even though the percentage of total casualties survived is similar for Model 1 and S(1,2), the breakdown of the results show the service strategies are different. An example of this is under Scenario 5 with 3 servers. In this case, Model 1 and S(1,2) have similar total results, but the percentage of types 1 and 2 survived shows that Model 1 generates mixed service policies as opposed to S(1,2) or S(2,1). Among the 5,000 simulation runs for the mentioned case, Model 1 generated the (unique) best solution 1,141 times, S(1,2) 929 times, and S(2,1) 818 times. Other times, at least two of the approaches generated the same best result. For the mentioned case, Model 1 generated S(1,2) 566 times, S(2,1) 52 times, and a mixed service policy 4,382 times. This shows while in a larger number of cases S(1,2) is preferred to S(2,1), still in majority of the runs the optimal solution is a more complex mixed policy.

2.5 More Than Two Casualty Types

Restricting our analysis to two casualty types provided insight into the structure of the optimal solution, and has some practical value ads discussed above. But there are often more casualty types, for instance, pediatric casualties have their own triage system, and would benefit from additional casualty types. Furthermore, Sacco et al. (2005) propose a triage system with 13 types. In this section, we expand our analysis beyond two casualty types.

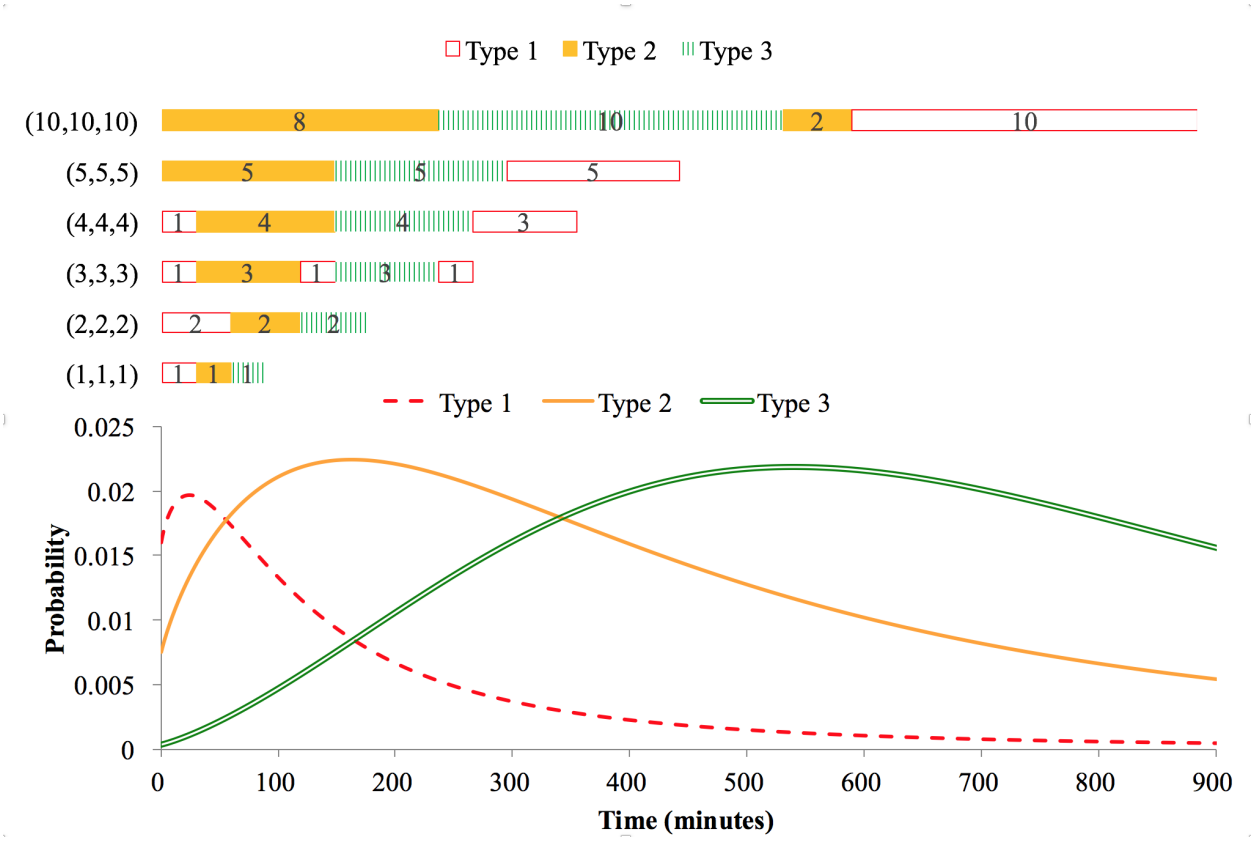


Figure 2.4: Survival probability difference and optimal solution for three casualty types

Figure 2.4 illustrates the survival probability function difference, f_{it}^{s}, for three casualty types. Casualty types in the figure follow the data in Table 2.2, with type 1 (red dashed curve) using the data from type 1 under Scenario 2, and types 2 (orange solid curve) and 3 (green dotted curve) using the data from types 1 and 2 under Scenario 5, respectively. Service times are 30 minutes for all casualty types. On the top portion of the figure, the optimal solutions for six cases with different number of casualties is depicted. The number of the casualties are shown as (n_1, n_2, n_3) next to the optimal service priority.

We observe that the structure of the optimal solution follows the pattern identified in Proposition 2 for two casualty types. Starting with $(1, 1, 1)$ case, the first t_s (service times are equal) is at time 55 for curves of type 1 and 2. Thus, the type 2 casualty is ordered around that point, with type 1 before and type 3 after it. Type 1 casualty is scheduled first as its f_{1t}^{s}

has a larger value than f^s_{3t}. For the second case, $(2, 2, 2)$, we have a similar optimal solution (i.e., $S(1, 2, 3)$). Two type 2 casualties are scheduled towards the right side of t_s of types 1 and 2 curves, as their difference is right-skewed (type 1 curve has a sharp decrease while type 2 has a gradual increase). As we increase number of the casualties, type 2 casualties get scheduled around the related $t_s = 55$, leaving insufficient room for all type 1 casualties to be served at the beginning. In case $(3, 3, 3)$, one type 1 is served after type 2s and before type 3s. $t_s = 164$ for curves of type 1 and 3, and type 3 casualties are scheduled around that point in $(3, 3, 3)$ case leaving room for one type 1 casualty before them. In $(5, 5, 5)$ case there is no place left for serving type 1 casualties before type 2 or type 3 and service becomes $S(2, 3, 1)$. Type 3 is the least critical type and its service cannot be interrupted by any of the other types, thus, service to type 1 is pushed to the end. In none of the mentioned five cases type 3 casualties are prioritized over type 2 as their corresponding $t_s = 341$, and both of these casualty types are served before that point in all cases. As stated in Proposition 2, service of a less critical casualty type is never interrupted by a more critical casualty type. In the last scenario, $(10, 10, 10)$, as we increase number of the casualty types, scheduling all type 3 casualties around $t_s = 341$ for curves of types 2 and 3 leaves insufficient room for initial service of all type 2 casualties, and thus, their service is broken into two parts, before type 3 casualties and after.

We can extend Algorithm 1 to solve the problem for multiple casualty types with equal service times and multiple servers. The extended approach is presented in Algorithm 3. The first change is introduction of LIST, which is the list of all possible casualty types assignments for every decision epoch and server. Here, instead of a single g_t value, we calculate the difference of survival probability curves for a specific casualty type and every other more critical casualty type, and then pick the maximum as the g_t for that item. For the most critical casualty type (type 1), this difference is always 0. After sorting the LIST based on the g_t value, we start picking items with the greatest g_t values. After each item is picked from LIST, we assign its casualty type to the same decision epoch and server in a results table, ORDER. Following each assignment, all the other items with the same decision epoch

and server are removed from the LIST, as no other assignment to the same server at that decision epoch is possible. We continue the same process until there is no other item in LIST. When LIST becomes empty, values in ORDER indicate type of the casualty to be served by the assigned server (corresponding column) at the specified time interval (corresponding row). This algorithm generates an optimal solution, as initially we enumerate all the possible assignments, and then pick the best at each iteration. Also, after each assignment, we only remove the other items for the same epoch and server, thus, each assignment is independent of other assignments for other epochs and servers. The complexity of this algorithm is the same as complexity of sorting the list, as the following steps are only iterating the list linearly. In the algorithm, EPOCH_j refers to the element j of the selected decision epoch, which are

time of service ($j = 1$), assigned server ($j = 2$), g_t value ($j = 3$), and type of casualty ($j = 4$).

Algorithm 3: Optimal algorithm for $|P|$ casualty types, $s_1 = s_2$, and m servers

```
1  begin
2      Create empty indexed-list LIST;
3      Create empty table ORDER;
4      k ← 1;
5      for t ← 1 to ⌈(n1+n2)/m⌉ do
6          for i ← 1 to m do
7              for j ← |P| to 1 do
8                  g_t ← max(f_jt − f_{j−1,t}, f_jt − f_{j−2,t}, ..., f_jt − f_{1,t});
9                  LIST[k] ← (t, i, g_t, j);
10                 k ← k + 1
11     sort LIST from largest to smallest (descending) value of the third element (g_t) for
         all items;
12     while |LIST| > 0 do
13         EPOCH ← Pick the first item in LIST (largest g_t value);
14         Add EPOCH_4 to row EPOCH_1 under column EPOCH_2 of ORDER;
15         Remove every element from LIST with time EPOCH_1 AND server EPOCH_2
```

Result: Values in ORDER indicate type of the casualty to be served by the assigned server (thier column) at each time interval (their row)

2.6 Conclusion and Future Steps

The medical community has identified the need to incorporate several factors such as scale of the disaster and availability of resources into the triage process. In this paper, we develop a mathematical model to show how the scale of the disaster and availability of the resources affects the outcome of the triage operation. Results from the analysis of our model provides

decision makers with optimal prioritization policies in order to maximize the expected number of survivors. We achieve this with minimal data requirements, specifically the number of casualties in each category, their service times, and the number of servers available. This information can be collected rapidly to generate the optimal triage plan.

Our findings verify the impact disaster-specific factors can have on the outcome of the triage process. This is in contrast with the current triage processes in practice such as START that have a static prioritization policy regardless of the scale of the disaster and available resources. We show although more critical-first policies similar to $S(1, 2)$ might be optimal in some cases, in many other, the reverse or some other mixed strategies perform significantly better. Our analysis allows us to gain some insight into the structure of the optimal policy, which could be the base for future studies and development of easy to implement, but precise policies. The model we present is a generalization of those found in the literature, which allows us to study problems under more practical assumptions, including multiple servers and arbitrary number of casualty types. A further enhancement is that our model readily accepts casualty specific service times, which adds realism. Despite these features, there are still some elements of a response to a mass-casualty indecent that should be considered in the future. We detail some of these in the following paragraphs.

Currently we assume that all casualties are served and there is no death before receiving service. While Frykberg and Tepas III (1988) find that there are small number of deaths among casualties before receiving service, casualty deaths need to be further investigated. We need to study how the possibility of death before receiving service affects the structure of the optimal solution. There has been research for the service order problem under the opposite assumption, namely that deaths occur before service (see Jacobson et al., 2012), but we feel this goes too far.

We also assume that all the casualties are available at the beginning of the relief effort, but in many cases, identification of all casualties might take hours. Stochastic casualty arrival can be incorporated into the model to address this issue and analyze how it affects the optimal

policies. Another assumption we made is having one disaster location. In some cases, there are multiple locations hit by a disaster and prioritization and transportation of the casualties should be done simultaneously in multiple locations.

Finally, hospital selection should be considered. Hospitals have limited capacity to treat casualties and different capabilities. These must be considered, else there is the possibility that the hospital becomes overwhelmed, which can lead to a reducing in the ability to treat casualties.

2.7 Appendix

Proof of Proposition 2. We divide the proof to two parts; first, we show the proposition holds for one server, and then, we show the works across multiple servers.

PART 1: Without loss of generality, consider a solution to the service order problem, having a partial schedule S starting at time t', which specifies serving one type 2 casualty, followed by m_1 type 1 casualties, and then a second type 2 casualty. We denote the total survival probabilities (i.e., expected number of survivors) generated by the schedule S, $TR(S)$, where $TR(S) = f_{2t'} + \sum_{r=0}^{m_1-1} f_{1t'+s_2+rs_1} + f_{2t'+s_2+m_1s_1}$. Under Assumption 2, we have $f_{1t}^{s_2} > f_{2t}^{s_1}$ for $t < t_s$, $f_{1t_s}^{s_2} = f_{2t_s}^{s_1}$, and $f_{1t}^{s_2} < f_{2t}^{s_1}$ for $t > t_s$. Thus, under Assumption 2 there are three possible cases for where t_s lies in regard to schedule S, and for each of these we show that schedule S, which interrupts the service of type 2 casualties, is suboptimal.

CASE 1: ($t_s > t' + 2s_2 + m_1s_1$, thus $f_{1t}^{s_2} > f_{2t}^{s_1}, \forall t = t', \cdots, t' + 2s_2 + m_1s_1$): Here we move the first type 2 casualty to the end, to form schedule S'. Now, we can calculate the difference in the total survival probabilities of schedules S and S' as follows:

$$
\begin{aligned}
TR(S') - TR(S) &= (f_{2t+m_1s_1} - f_{2t}) + \sum_{i=1}^{m_1} \left(f_{1t+(i-1)s_1} - f_{1t+s_2+(i-1)s_1}\right) \\
&= (f_{2t+m_1s_1} - f_{2t}) - \sum_{i=1}^{m_1} \left(f_{1t+s_2+(i-1)s_1} - f_{1t+(i-1)s_1}\right) \\
&= \left(f_{2t+m_1s_1} + \sum_{i=1}^{m_1-1} \left(f_{2t+is_1} - f_{2t+is_1}\right) - f_{2t}\right) \\
&\quad - \sum_{i=1}^{m_1} \left(f_{1t+s_2+(i-1)s_1} - f_{1t+(i-1)s_1}\right) \\
&= \sum_{i=1}^{m_1} \left(f_{2t+is_1} - f_{2t+(i-1)s_1}\right) - \sum_{i=1}^{m_1} \left(f_{1t+s_2+(i-1)s_1} - f_{1t+(i-1)s_1}\right) \\
TR(S') - TR(S) &= -\sum_{i=1}^{m_1} f_{1t+(i-1)s_1}^{s_2} + \sum_{i=1}^{m_1} f_{2t+(i-1)s_1}^{s_1} \qquad (2.20)
\end{aligned}
$$

There are m_1 terms in each summation of the equation (2.20). We now compare terms in the summations with each other one by one. Given the assumption for this case ($f_{1t}^{s_2} > f_{2t}^{s_1}$), we have $f_{1t}^{s_2} > f_{2t}^{s_1}$, $f_{1t+s_1}^{s_2} > f_{2t+s_1}^{s_1}$, $f_{1t+2s_1}^{s_2} > f_{2t+2s_1}^{s_1}$, ..., and $f_{1t+(m_1-1)s_1}^{s_2} > f_{2t+(m_1-1)s_1}^{s_1}$. Thus, we conclude that $\sum_{i=1}^{m_1} f_{1t+(i-1)s_1}^{s_2} > \sum_{i=1}^{m_1} f_{2t+(i-1)s_1}^{s_1}$, and $TR(S') > TR(S)$. As a result, when $f_{1t}^{s_2} > f_{2t}^{s_1}$, we are always better off serving type 1 casualties first, and then, type 2 casualties.

CASE 2: ($t_s < t'$, thus $f_{1t}^{s_2} < f_{2t}^{s_1}, \forall t = t', \cdots, t' + 2s_2 + m_1 s_1$): In this case we move the second type 2 casualty to the beginning after the first one, to form schedule S''. Now, we can calculate the difference in the total survival probabilities of schedules S and S'' as follows:

$$\begin{aligned}
TR(S'') - TR(S) &= (f_{2t+s_2} - f_{2t+m_1 s_1+s_2}) + \sum_{i=1}^{m_1} \left(f_{1t+2s_2+(i-1)s_1} - f_{1t+s_2+(i-1)s_1}\right) \\
&= \sum_{i=1}^{m_1} \left(f_{1t+2s_2+(i-1)s_1} - f_{1t+s_2+(i-1)s_1}\right) - (f_{2t+m_1 s_1+s_2} - f_{2t+s_2}) \\
&= \sum_{i=1}^{m_1} \left(f_{1t+2s_2+(i-1)s_1} - f_{1t+s_2+(i-1)s_1}\right) \\
&\quad - \left(f_{2t+m_1 s_1+s_2} \sum_{i=1}^{m_1-1} (f_{2is_1+s_2} - f_{2is_1+s_2}) - f_{2t+s_2}\right) \\
&= \sum_{i=1}^{m_1} \left(f_{1t+2s_2+(i-1)s_1} - f_{1t+s_2+(i-1)s_1}\right) \\
&\quad - \sum_{i=1}^{m_1} \left(f_{2t+s_1+(i-1)s_1+s_2} - f_{2t+(i-1)s_1+s_2}\right) \\
TR(S'') - TR(S) &= -\sum_{i=1}^{m_1} f_{1t+s_2+(i-1)s_1}^{s_2} + \sum_{i=1}^{m_1} f_{2t+(i-1)s_1+s_2}^{s_1} \qquad (2.21)
\end{aligned}$$

Equation (2.21) has a similar structure to equation (2.20) and using a similar comparison, we conclude that $TR(S'') > TR(S)$. Therefore, when $f_{2t}^{s_1} > f_{1t}^{s_2}$, we are always better off serving type 2 casualties first, and then, type 1 casualties.

CASE 3: ($t' < t_s < t' + 2s_2 + m_1 s_1$, thus $f_{1t}^{s_2} > f_{2t}^{s_1}, \forall t = t', \cdots, t_s$ and $f_{1t}^{s_2} < f_{2t}^{s_1}, \forall t = t_s, \cdots, t' + 2s_2 + m_1 s_1$): To study this case, we break it into two time intervals, $[t', t_s]$ and $[t_s, t + m_1 s_1 + 2s_2]$. In the first interval we have $f_{2t}^{s_1} < f_{1t}^{s_2}$. We showed in Case 1 that under this setting, we are better off serving type 1 casualties first and then type 2 casualties. In the second interval, we have $f_{2t}^{s_1} > f_{1t}^{s_2}$ as in Case 2, where we showed it is better to serving type 2 casualties first, and then type 1 casualties. Combining the results from these two intervals, we have a portion of type 1 casualties served first, then all type 2 casualties, and finally rest of the type 1 casualties.

Given the three possible cases, it is optimal not to interrupting service of type 2 casualties for a server.

PART 2: Without loss of generality, assume we have a partial schedule S with two servers starting at time t', where server 1 serves two type 1 casualties and server 2 serves two type 2. Same as PART 1, there are three cases in regard to where t_s lies in regard to schedule S. For each of the three cases we show that schedule S which is not balanced among servers in terms of service to difference casualty types, is suboptimal.

CASE 1: ($t_s > \max(2s_1, 2s_2)$, thus $f_{2t}^{\delta} < f_{1t}^{\delta'}, \forall t = t', t' + s_1, t' + s_2$ and $\delta, \delta' > 0$): In this case, we switch the second type 1 casualty in server 1 with the first type 2 casualty in server 2 to form schedule S'. The difference in total survival probabilities of schedules S' and S is as follows:

$$\begin{aligned}
TR(S') - TR(S) &= f_{1t} + f_{2t+s_1} + f_{1t} + f_{2t+s_1} - f_{1t} - f_{1t+s_1} - f_{2t} - f_{2t+s_2} \\
&= f_{2t+s_1} + f_{1t} + f_{2t+s_1} - f_{1t+s_1} - f_{2t} - f_{2t+s_2} \\
&= f_{2t+s_1} + f_{1t} + f_{2t+s_1} - f_{1t+s_1} - f_{2t} - f_{2t+s_2} \pm f_{1t+s_2} \\
&= (f_{1t}^{s_2} - f_{2t}^{s_1}) + \left(f_{1t+s_2}^{s_1-s_2} - f_{2t+s_2}^{s_1-s_2}\right)
\end{aligned}$$

In this case, $f_{2t}^{s_2} < f_{1t}^{s_1}, \forall t = t', t'+s_1, t'+s_2$. Thus, both terms $(f_{1t}^{s_2} - f_{2t}^{s_1})$ and $\left(f_{1t+s_2}^{s_1-s_2} - f_{2t+s_2}^{s_1-s_2}\right)$ are positive. As a result, $TR(S') - TR(S) > 0$, which indicates unbalanced service among servers cannot be optimal.

CASE 2: ($t_s < t'$, thus $f_{1t}^{\delta} < f_{2t}^{\delta'}, \forall t = t', t'+s_1, t'+s_2$ and $\delta, \delta' > 0$):In this case, we switch the first type 1 casualty in server 1 with the second type 2 casualty in server 2 to form schedule S''. The difference in total survival probabilities of schedules S'' and S is as follows:

$$\begin{aligned} TR(S'') - TR(S) &= f_{2t} + f_{1t+s_2} + f_{2t} + f_{1t+s_2} - f_{1t} - f_{1t+s_1} - f_{2t} - f_{2t+s_2} \\ &= f_{1t+s_2} + f_{2t} + f_{1t+s_2} - f_{1t} - f_{1t+s_1} - f_{2t+s_2} \\ &= f_{1t+s_2} + f_{2t} + f_{1t+s_2} - f_{1t} - f_{1t+s_1} - f_{2t+s_2} \pm f_{2t+s_1} \\ &= (f_{2t}^{s_1} - f_{1t}^{s_2}) + \left(f_{2t+s_2}^{s_1-s_2} - f_{1t+s_2}^{s_1-s_2}\right) \end{aligned}$$

Based on the assumption in this case ($f_{2t}^{s_2} < f_{1t}^{s_1}, \forall t = t', t'+s_1, t'+s_2$), both terms above $(f_{2t}^{s_1} - f_{1t}^{s_2})$ and $\left(f_{2t+s_2}^{s_1-s_2} - f_{1t+s_2}^{s_1-s_2}\right)$ are positive. Thus, we are better off balancing out service on both servers.

CASE 3: ($t' < t_s < \max(2s_1, 2s_2)$, thus $f_{1t}^{\delta} > f_{2t}^{\delta'}, \forall t = t', \cdots, t_s$ and $f_{1t}^{\delta} < f_{2t}^{\delta'}, \forall t > t_s$): We break this case into two time intervals, $[t, t_s]$ and $[t_s, n_1 s_1 + n_2 s_2]$. In the first interval we have $f_{2t}^{s_1} < f_{1t}^{s_2}$ similar to CASE 1, in which we showed that we are better off balancing service among two servers by arranging service to type 2 casualties towards the end. In the second interval($[t_s, n_1 s_1 + n_2 s_2]$), we have $f_{1t}^{\delta} < f_{2t}^{\delta'}, \forall t > t_s$ similar to CASE 2, in which we showed that we are better off balancing servers by arranging type 2 casualties at the beginning for each server. Combining the results from these two intervals, we have balanced servers serving type 2 casualties around (or as close as possible to) t_s.

The results from PART 2 can be extended to more than two servers by comparing two servers

at a time. We can conclude that under Assumption 2, in the optimal solution to Model 1, service to type 2 casualties is not interrupted. $\square$

Chapter 3

Service priority for mass-casualty incident response

Abstract

Mass casualty incidents overwhelm medical resources and casualties have to be prioritized in order to receive service. Triage is the process of categorizing casualties and most triage methods come with an implied service prioritization order, starting with the most critical casualties, regardless of the resources available. In this paper we study and analyze multiple approaches for modeling the service priority problem. We focus on how delay in service deteriorates the casualties' condition to potentially beyond survival, and how this transition affects the outcome. The data used in the study and its effects on the results is discussed in details. We study properties of the optimal policies under special cases and develop a heuristic approach. Finally, a numerical analysis and performance comparison with other methods in the literature is provided.

3.1 Introduction

Mass-casualty incidents (MCI) overwhelm medical resources (e.g., ambulances) due to a sudden increase in demand in a short window of time. Limited resources should be assigned as efficiently as possible among casualties. In fact, management of resources is identified as the most critical factor in MCI response efforts (Antommaria et al., 2011). Following an MCI, upon arrival, the first responders triage casualties into several categories based on the severity of their conditions. There are several MCI triage methods, and each one is developed with the goal of utilizing available resources in the most efficient form. More well-known triage methods include START (Simple Triage and Rapid Treatment Method), Homebush, Triage Sieve, Sacco Triage Method (STM), and CESIRA (Lerner et al., 2008). These are not to be confused with hospital emergency department triage methods (see Iserson and Moskop, 2007, for details). START, which is the most common triage method in the US, has four color-coded categories: immediate (red), delayed (yellow), minor (green), and expectant (black) (Cone and MacMillan, 2005). There are also several pediatric triage methods developed for

addressing children needs, most well-known of which is JumpSTART that is evolved from START for children 1 to 8 years old (Jenkins et al., 2008).

Most triage methods come with an implied service order for serving casualties based on their assigned category. The implied service order is prioritizing casualty categories from the most critical to the least critical (e.g., red, yellow, and then green in case of START). This prioritization does not account for the available resources, the scale of the incident, or the casualty mix. A review of the literature reveals that effectiveness and performance of the START or any of the other common triage methods have not been scientifically evaluated (Jenkins et al., 2008; Lerner et al., 2008). The measure of triage effectiveness we use in this paper is the survival rate among the casualties that are treatable (e.g., those in the red, yellow, and green START categories) in the hours to days following the incident. In this paper we identify and analyze the best policies to maximize survival rate of an MCI considering resources and number of the casualties.

For modeling purposes, there are two main simplifying assumption about when and where deaths occur in the operations research literature, the *on-site death* assumption where casualties could only die at the location of the MCI, before service (i.e., transport to a hospital) and therefore do not use scarce resources, or the *off-site death* assumption where casualties could only die at the hospital, after scarce emergency resources have been expended on them. Under this assumption casualties deteriorate while waiting for service, but the actual outcome is not observed until after the service. Based on our reading of the medical and emergency management literature, of these two assumptions, the *off-site death* assumption is more realistic. For instance, Frykberg and Tepas III (1988) study 220 terrorist bombing incidents with 2,934 casualties surviving the attack and waiting for service. They recorded 40 subsequent deaths among the 2,934 survivors, of which only one occurred prior to receiving service. In a similar study of the 2005 London bombings, Aylwin et al. (2007) divide deaths into prompt deaths and potentially preventable subsequent deaths. Although what they refer to as a prompt death sounds similar to on-site death, they are different, as prompt death is assumed to be unpreventable, regardless of service order and are better thought of

as the expectant (black) category from START. When considering survival probability, we only consider casualties whose death is preventable by a timely delivery of care. In another study, Sacco et al. (2005) develop a triage system, STM, in which they assume there are no on-site deaths, and the off-site survival probabilities (presented as service time dependent survival probabilities for each category of casualty) are generated from historical data.

Of course considering casualty deaths both on-site and off-site is probably most appropriate, but it is more difficult to model analytically because of the complexities; there is no clear dividing line between on-site and off-site deaths, and deaths that occur in a hospital under one resource level could occur on-site when less resources are available. Furthermore, some on-site deaths can be the result of misclassification; rapid classification of casualties is challenging and subject to inaccuracy. There are two types of misclassification: assignment of a more-critical casualty to a less-critical type, *undertriage*, and assignment of a less-critical casualty to a more-critical type, *overtriage*. Frykberg (2002) state that undertriage could lead to preventable deaths and it should be avoided in both daily and mass-casualty triage operations, but state no reported case of undertriage exists in the past bombing attacks. In dealing with daily trauma patients, overtriage occurs at a rate of around 50% (Kreis Jr et al., 1988), which can also be necessary to avoid preventable deaths by reducing undertriage as much as possible (Frykberg, 2002). On the other hand, Frykberg and Tepas III (1988) conclude overtriage could be as deadly as undertriage in a mass-casualty incident. Overtriage of less-critical casualties could delay service for more-critical casualties and potentially threaten their survival. In contrast, Hupert et al. (2007) find that overtriage could have mixed effects on outcome of triage operation through simulation analysis. In this paper, we study the optimal triage service order problem, as well as the cases under the on-site, off-site assumptions (triage misclassification is beyond the scope of this paper), and explore how these assumptions would affect the performance of our solutions and others developed in the literature.

Here we briefly review other papers that study this problem in the operations research literature and provide more details in later sections as appropriate. Sacco et al. (2005,

2007) study the triage problem under off-site death assumption, and they maximize sum of decreasing survival probabilities for all casualties by determining their service time. Dean and Nair (2014) develop a model similar to Sacco et al. (2005), with addition of hospital bed capacity and resource reusability. Kamali et al. (2016); Mills et al. (2011, 2013) similarly study the resource-based triage problem under the off-site death assumption, using a model that we discuss in more details later. Argon et al. (2008, 2011); Jacobson et al. (2012) study the casualty priority assignment problem under the on-site death assumption. The model in Jacobson et al. (2012), which we discuss in more detail later, utilizes a casualty type based reward, which can be thought of as the probability of off-site deaths, but it is constant and independent of the delay until service. Li and Glazebrook (2011) study the triage problem considering imperfect classification of casualties. They take a Bayesian approach in dealing with uncertainties and develop heuristics approaches for the problem. These papers all show that resource levels, relative to the number of casualties, can affect the optimal service order, but none of which consider both off-site and on-site deaths in their model.

Our main contributions in this study include: **1)** develop a general model for the service order problem, and analyzing the transformation to models under both off-site and on-site death assumptions; **2)** discuss and analyze the data and parameters used in this problem in details; **3)** provide insight into the structure of the optimal policies and analytical comparison of our models and others in the literature; and, **4)** analyze numerical results from our models and other in the literature under different assumptions and discuss the effect of incorporating resources and casualties mix on the outcome of the triage process.

The remainder of this paper is organized as follows. In §2, we introduce our mathematical models and compare them with the other models developed in the literature. In §3, we define the parameters used in this study, and analyze the structure of the data in depth. In §4, we propose a two-casualty type simplification to the model and discuss the results. In §5 and §6, we analyze the proposed models analytically and numerically, respectively, and extend the comparison to the related models in the literature. In §7, we briefly discuss the case with multiple servers, and finally in §8, concluding remarks are stated.

3.2 Models Description

In this section, we first describe the service order problem. Consider a mass casualty incident where casualties are categorized into P types, with type 1 being the most critical casualty type (e.g., immediate in START, with shortest expected lifetime), and type P the least critical type (e.g., minor in START, with longest expected lifetime). In the casualty classification we only consider casualties that are expected to survive, assuming there is no shortage of the resources. For instance, we do not consider START's expectant category, which has no or very low chance of survival regardless of the resources. The objective is to determine a service order plan that maximizes the number of survivors, given the number of casualties, denoted n_i for type $i = 1, \ldots, P$, and survival probability, f_{it}, for each casualty type i for delaying service until time t. We define the survival probability (actual) in terms of the casualties that can survive, if they receive service in a timely manner (See, for instance, Sacco et al., 2007, which provides survival curves, from which survival probabilities can be calculated.). As in Jacobson et al. (2012); Mills et al. (2013), we assume all casualties are available to serve at time zero, there is one server (e.g., an ambulance) that works in a non-preemptive manner, and an expected service time $s_i,\ i = 1, \ldots, P$ for each casualty type, which encompasses loading and unloading time, any care required to stabilize the casualty, and the round trip travel time to an appropriate hospital. The time horizon required for the disaster response is divided into T equal time intervals.

To provide the most similarities to the real-world practice, we assume a casualty's death (e.g., not surviving or eventual death similar to START's expectant category, which in other words occurs when the casualty becomes non-treatable) can be observed both on-site, prior to receiving service, and off-site, after resources are occupied to serve the casualty. If death (lack of survivability) is observed on-site, the next casualty in the queue is served (assuming there is at least one casualty left). On the other hand, if a casualty dies off-site, resources are occupied, while no reward is collected. This situation resembles a casualty in the START's immediate or delayed category when they have waited for some time on the

disaster location without receiving service, and their condition degrades continuously. After some point, depending on their condition, casualty is likely to become non-survivable and their category changes to expectant, that is, regardless of the resources, there is no chance for their survival. At that point, there is a chance that emergency personnel can observe the casualty's condition and provide service to another casualty with a higher chance of survival, instead.

To model this problem, we define a probability for observing that a casualty of type i's condition has degraded past survival (i.e., switch to START expectant category) by time t, denoted by O_{it} for $i = 1, \ldots, P$ and $t = 1, \ldots, T$. As casualties wait longer before receiving service, their condition degrades and it becomes more apparent to the emergency personnel if they will not survive regardless of the resources. Hence, O_{it} increases in time and approaches to 1 as $t \to \infty$, that is, casualty's death will be observed. We formulate the problem as a dynamic programming (DP) model. The action space has a size equal to size of the casualty types at each decision epoch, in each of which a casualty is removed from the system. Removal from the system can be of 3 forms based on the survival and death observation probabilities; service with collecting reward (i.e., survival), service without collecting reward (i.e., off-site death), and discharge from the system with no service or reward (i.e., on-site death). The modeling parameters and the developed model are as follows:

Sets and Parameters:

- n_i number of casualties of type i at the beginning of time interval 1, $\forall i = 1, \ldots, P$
- s_i number of time intervals it takes to serve a casualty of type i, $\forall i = 1, \ldots, P$
- T number of time intervals in the time horizon ($\sum_{i=1}^{P} s_i n_i$)
- α_i reward collected by serving a casualty of type i, $\forall i = 1, \ldots, P$ when survivable
- f_{it} survival probability associated with delaying service for a casualty of type i until time interval t, $\forall i = 1, \ldots, P, t = 1, \ldots, T$
- O_{it} probability of observing death for a casualty of type i at time interval t, $\forall i = 1, \ldots, P, t = 1, \ldots, T$

Q vector of number of remaining casualties of each type to be served, $(q_1, q_2, \ldots, q_P)$

$$\begin{aligned}\textbf{Model 1: } V(Q,t) = \max_{i=1,\ldots,P}\{ & f_{it}(\mathbb{I}_{\{q_i>0\}}\alpha_i + V(Q : q_i - 1, t + s_i)) \\ & + (1 - f_{it})((1 - O_{it})V(Q : q_i - 1, t + s_i) + O_{it}V(Q : q_i - 1, t))\},\end{aligned} \tag{3.1}$$

where $V(Q,t)$ is the value function of the state variables, Q and t. Terminating conditions are when $q_i = 0$ for all $i = 1, \ldots, P$, or $q_i < 0$, for any $i = 1, \ldots, P$, in which $V(Q,t) = 0$. $\mathbb{I}_A$ is an indicator function that is 1 if A is true, and 0 otherwise. Decision epochs in this model are time one and service completion times. At each epoch the decision is which casualty type to serve from those having $q_i > 0$. To find the optimal service policy, we seek $V(Q,1)$, where $Q = (n_1, n_2, \ldots, n_P)$. There are two recursive terms in the formulation; $V(Q : q_i - 1, t + s_i)$ refers to serving a current casualty of type i and moving to the next state at time period $t + s_i$, and $V(Q : q_i - 1, t)$ indicates eliminating a type i casualty from the system without service. When serving a type i casualty, we only collect the reward (α_i) if the casualty's lifetime is not expired (i.e., with probability f_{it} at time t). Next, we present some insights from our Model 1 and related models under simplifying assumptions used in the literature, namely off-site death assumption (i.e., having no information about casualty's death and only having off-site deaths, $O_{it} = 0$), and on-site death assumption (i.e., having perfect information of casualty's death and only having on-site deaths, $O_{it} = 1$).

3.2.1 Insights

Having the survival and death observation probabilities, using the developed Model 1 we can derive on-site and off-site survival probabilities. There are two possible cases that a casualty is observed to survive on-site: first, when the casualty actually survives, with probability f_{it},

and second, when the casualty's condition is deteriorated into expectant category, but it is not observed, with probability $(1 - f_{it})(1 - O_{it})$. In the latter case, although casualty is not going to survive eventually, they are observed to survive on-site and be served. Thus, the on-site survival probability is:

$$P(\text{on-site survival}) = f_{it} + (1 - f_{it})(1 - O_{it}), \tag{3.2}$$

where a casualty is perceived survived on the disaster location, and is transported to care facility. On the other hand, among the casualties surviving on-site (and being served), only the non-expectant portion are going to survive off-site. Thus, the off-site survival probability is:

$$P(\text{off-site survival}) = \frac{f_{it}}{f_{it} + (1 - f_{it})(1 - O_{it})}, \tag{3.3}$$

where a reward is collected for the service to the casualty. This is a conditional probability of actual survival, given the casualty is survived on-site. The complement to the off-site survival, is off-site death, which can also be considered as wasted service. That is, when a casualty is served, which is not going to survive, their service is wasted. Therefore, we can calculate the portion (probability) of the wasted service as follows:

$$P(\text{wasted service}) = \frac{(1 - f_{it})(1 - O_{it})}{f_{it} + (1 - f_{it})(1 - O_{it})}. \tag{3.4}$$

Off-site Death Assumption ($O_{it} = 0$)

Next, we present Model 2, a DP that determines the optimal strategy to maximize survival probability for the service ordering problem under the *off-site death* assumption. Under this assumption, we presume that we have no information about the degradation in casualty's condition (i.e., the O_{it} distribution having mean value of ∞ and thus having no on-site

deaths) and all casualties are served. In other words, we assume no deaths are observed on-site prior to receiving service, $O_{it} = 0$, for all $i = 1, \ldots, P$ and $t = 1, \ldots, T$, which cancels out the term $O_{it}V(Q : q_i - 1, t)$ in Model 1 and simplifies it to the following (both $V(Q : q_i - 1, t + s_i)$ terms are merged):

$$\textbf{Model 2: } V(Q, t) = \max_{i=1,\ldots,P} \{f_{it}\mathbb{I}_{\{q_i>0\}}\alpha_i + V(Q : q_i - 1, t + s_i)\}, \tag{3.5}$$

where base cases are similar to Model 1. The decision epochs are time one and service completion times. In this model, survival probability decreases the longer casualties remain at the disaster location, and once a casualty begins transport to the hospital, the survival probability is realized, and it determines whether the reward is collected or not. On-site survival probability (3.2) becomes equal to 1 under off-site death assumption, as no on-site deaths are observed and everyone is served, and off-site survival probability (3.3) becomes equal to actual survival probability, f_{it}. Also, probability of wasted service (3.4) becomes $(1 - f_{it})$. Few other lines of research study the service order problem under off-site death assumption. Here, we briefly mention them and compare them to Model 2.

Kamali et al. (2016) develop a linear programming model to address the service order problem for multiple casualty types and servers under off-site death assumption. We show in Proposition 3 that Model 2 provides the optimal solution to their model with one server, and also the DP provides a more complex solution, much of which can be considered extraneous under the *off-site death* assumption, but as we show in §6, this additional solution complexity can be useful for adapting the solution to a general setting. While this is not the case with Kamali et al. (2016) model, a binary programming also has some advantages. It is useful for comparison to other models in the literature, and it can be readily modified to an integer program to include multiple servers and other added complexities, which would tend to make the DP intractable since DPs generate the results for the complete state space. Kamali et al. (2016) show that their model has a totally unimodular constraint coefficient matrix and it produces integer results when integrality constraints on its variables are re-

laxed. Hence, their model can be solved for large cases in a tractable time, even with added complexities.

Proposition 3 *Model 2 provides an optimal solution for the Mills et al. (2013) fluid formulation, Sacco et al. (2007), and Kamali et al. (2016) models with one server.*

Proof. See Appendix.

Mills et al. (2013) propose a fluid formulation to study the service order problem under off-site death assumption. Their model maximizes the survival probability under the assumption that service times are equal for both casualty types (e.g., service time is the transportation time from the disaster location to the receiving hospital). They use their proposed formulation to gain some insight into the optimal policies, and propose two heuristics to address the problem. We discuss their heuristic approaches in our numerical analysis section. Sacco et al. (2005, 2007) also propose a linear programming model under off-site death assumption that can address multiple casualty types, but with equal service times limitation. Similar to Kamali et al. (2016), they consider variability in number of the resources (e.g., ambulances) through time, although unlike Kamali et al. (2016), they do not consider reusing resources that become available after their initial service ends. Results from Model 2 can be represented in a vector format as it assumes all casualties are served. Therefore, there is no uncertainty in time when number of the remaining casualties and service times are given. Vector-based solution contains $\sum_{i \in P} n_i$ elements, each of which indicates a service type. On the other hand, Model 1 generates a service graph that has a solution for all possible states (not necessarily optimal), as there is a probability to observe casualty's death on-site and refrain service. This is a major advantage to Model 2 in terms of tractability. Next, we study the problem under on-site death assumption and briefly review the related works in the literature.

On-site Death Assumption ($O_{it} = 1$)

Under on-site death assumption, we assume we have perfect information about a casualty's survival chance. This means that service is only provided to casualties, if they are going to survive (i.e., there is no on-site deaths). In this case, lifetime of each casualty is observed before service, and if the lifetime is not expired, service is provided and resources become occupied. Otherwise, the casualty is removed from the system without receiving service. Model 3 formulates the problem under this assumption by setting $O_{it} = 1$, which removes the term $(1 - O_{it})V(Q : q_i - 1, t + s_1)$ from Model 1.

$$\begin{aligned} \textbf{Model 3: } V(Q,t) = \max_{i=1,\dots,P} \{ & f_{it}(\mathbb{I}_{\{q_i>0\}}\alpha_i + V(Q : q_i - 1, t + s_i)) \\ & + (1 - f_{it})V(Q : q_i - 1, t)\} \end{aligned} \tag{3.6}$$

Same as Model 1, there are P choices at every decision epoch; Serving a type i casualty for all $i = 1, \dots P$. Each choice can stochastically result in two other states. First is when a casualty of the selected type is served, reward is collected, and time is advanced (i.e., resources are occupied), and the other is when the casualty is removed from the system without service with a probability of mortality (complement of the survival probability) at the observation time. Terminating states are similar to that of Model 1 and to find optimal solution, we seek $V(Q, 0)$ for $Q = (n_1, \dots, n_P)$. Under this assumption, probability of on-site survival (3.2) becomes equal to actual survival probability, f_{it}, as we assume we have perfect information about eventual outcome of the service, and thus, probability of off-site survival (3.3) becomes 1. Aslo, wasted service probability (3.4) is 0 under this assumption, as there are no off-site deaths.

To the best of our knowledge, Jacobson et al. (2012) is the only study in the literature that examines the casualty prioritization problem under the *on-site death* assumption. That

paper uses a Markovian model using exponentially distributed lifetimes and service times with two casualty types and one server. Casualties die on-site if their service does not begin by the end of their lifetimes, but once service begins, it is assumed that the casualty survives, and a constant reward is collected based on the casualty's type. The objective is to maximize the expected total reward by determining a policy (π^*) with the best service order. Jacobson et al. (2012) proposes four different heuristic approaches to solve this problem, of which we discuss two in our numerical analysis section. Next, we discuss the data used in our analysis of the triage problem.

3.3 Data

There are three sets of parameters used in our models; survival probability, service time, and probability of observing death. Except for service times that we assume to be deterministic, our models are able to use any discretized distribution for the other two parameter sets. In this section we briefly discuss the data and distributions used in the literature for each set of the mentioned parameters.

First, we define the following to refer to the change in survival probabilities through time.

Definition 2 $f_{it}^{\delta} \equiv f_{it+\delta} - f_{it}$.

f_{it}^{δ} denotes the decrease in the survival probability, f_{it}, starting at time t for δ time intervals. Given this definition, we make the following assumption about the survival probabilities, which is a discretized generalization of the assumption made by Mills et al. (2013), including rewards and unequal service times:

Assumption 3 *There exists a time* t_m *such that* $\alpha_1 f_{1t}^{\delta} > \alpha_2 f_{2t}^{\delta} > 0$ *for all* $t < t_m$, $\alpha_1 f_{1t_m}^{\delta} = \alpha_2 f_{2t_m}^{\delta}$, *and* $\alpha_2 f_{2t}^{\delta} > \alpha_1 f_{1t}^{\delta} > 0$ *for all* $t > t_m$.

Assumption 3 implies that there is a point, t_m, at which $\alpha_2 f_{2t}$ (i.e., the reward collected for serving a casualty of type 2, if served at time t) starts to decrease faster than $\alpha_1 f_{1t}$. If $t_m < 0$, then $\alpha_2 f_{2t}$ decreases with faster rate than f_{1t} for all time intervals studied, while if $t_m > T$, then $\alpha_1 f_{1t}$ grows faster than $\alpha_1 f_{2t}$ for all time intervals studied. In any other case, the growth rate switches at time $t_m \in [0, T]$. Mills et al. (2013) uses a simpler version of assumption (with no rewards and equal service times) in their analysis, and given that their problem definition has $s_1 = s_2$, this is equivalent to $\alpha_1 f_{1t_m} - \alpha_2 f_{2t_m}$ having a unique maximum at $t_m \in [0, T]$. Thus, $\alpha_1 f_{1t_m} - \alpha_2 f_{2t_m}$ is increasing for $t < t_m$, and is decreasing for $t > t_m$. This assumption holds for exponential survival probabilities and deterministic service times that we use throughout this paper.

Proposition 4 *Exponentially distributed survival probabilities with deterministic service times and constant rewards adhere to Assumption 3.*

Proof. See Appendix.

To estimate a casualty's survival probabilities, the exponential distribution has been used in the literature, but no previous study has shown that exponential distribution is a good distribution for lifetimes in an emergency response effort. Our models can use any distribution for casualty's survival probabilities, but here we use exponential distribution, as it has also been used by Jacobson et al. (2012) and it adheres to the Assumption 3. In addition to exponential distribution, Mills et al. (2013) develop a scaled log-logistics distribution-based Function (3.7) based on the analysis done on historical data by Sacco et al. (2005). This function has three parameters, of which $\beta_{0,i}$ is the scalar and $\beta_{1,i}$ is the mean. Throughout this paper we perform our analysis with exponential distribution, as it adheres to Assumption 3, while for Function (3.7), this assumption only holds with a certain parameters. For instance, for $(\beta_{0,1}, \beta_{1,1}, \beta_{2,1}) = (0.39, 53, 4.01)$, $(\beta_{0,1}, \beta_{1,1}, \beta_{2,1}) = (0.57, 61, 2.03)$, and $\alpha_1 = \alpha_2 = 1$, $\alpha_2 f_{2t}^{\delta} - \alpha_1 f_{1t}^{\delta}$ has multiple extremums, as opposed to one (i.e., t_m) required

by the Assumption 3. In addition, the parameters proposed by Mills et al. (2013) for Function 2.19 unnecessarily decrease initial survival probabilities (see Kamali et al., 2016, for detailed discussion)

$$f_i(t) = \frac{\beta_{0,i}}{(t/\beta_{1,i})^{\beta_{2,i}} + 1}, \quad \text{for all } i \in P \tag{3.7}$$

We also use exponential distribution for O_{it}, in order to capture the increasing nature of the probability of observing death prior to receiving service, with mean d_i for casualty type $i \in P$ (i.e., $1 - e^{-\frac{t}{d_i}}$). In addition, exponential distribution shows the increasing speed in which emergency personnel will be able to identify deceased (expectant) casualties, as their condition degrades. Figure 3.1 illustrates sample survival (f_{it}), observability (O_{it}), earlier defined on-site survival (3.2), off-site survival (3.3), and wasted service (3.4) using exponentially distributed actual survival and observation probabilities with means 200 and 300, respectively.

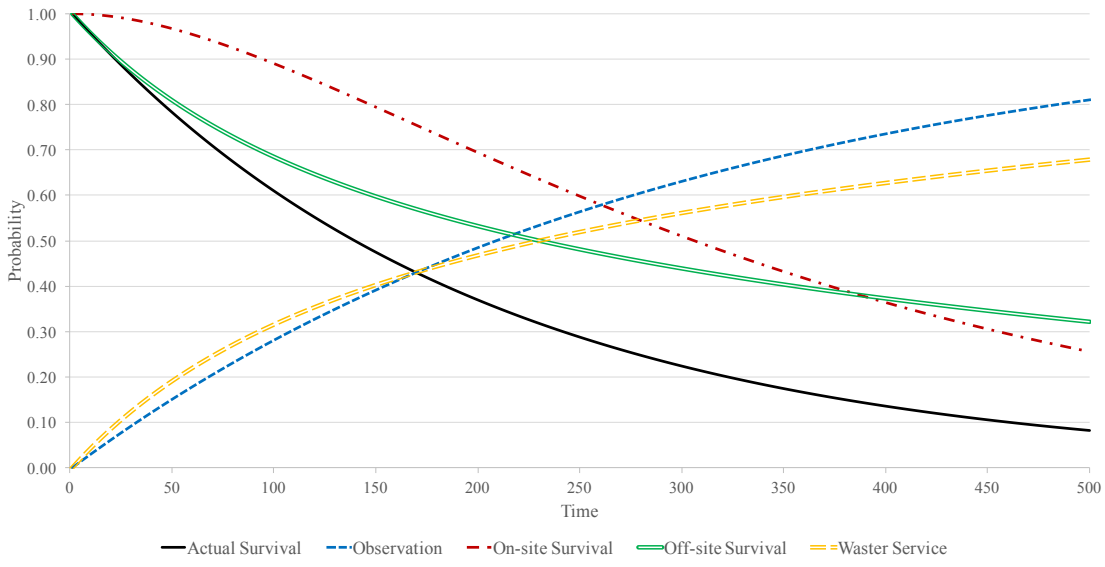


Figure 3.1: Survival with mean 200, observation with mean 300, on-site survival, off-site survival, and wasted service probabilities

Proposition 5 *The probability of observing death on-site has a mean of* $l_i + d_i - \frac{l_i d_i}{l_i + d_i}$, *when*

the survival and observation probabilities follow exponential distribution with means l_i and d_i for a casualty type i, respectively.

PROOF. See Appendix.

For service times (mainly patient transportation time), log-normal distribution is shown to be a good fit (Ingolfsson et al., 2008). Our developed models use constant service times (same as (Sacco et al., 2007; Mills et al., 2013; Kamali et al., 2016)), but Jacobson et al. (2012) use exponential distribution for analytical purposes. Exponential service times are biased towards smaller values, which contradicts the nature of MCIs with scarce resources. Figure 3.2 shows exponential and log-normal service times with mean of 30 minutes. The mean is highlighted with a vertical line in the figure. Later in our numerical analysis, we show that using constant service times does not affect the quality of the results from our models significantly.

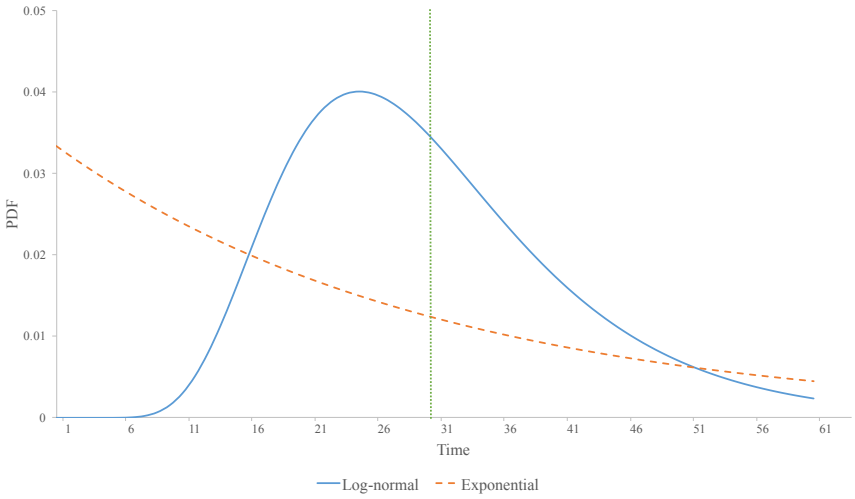


Figure 3.2: Exponential and log-normal service times with mean 30 minutes and log-normal standard deviation of 12 minutes

In our analysis and observations throughout this paper, across different scenarios, we keep the expected lifetimes the same and only vary service times. This has a higher resemblance to the reality, compared to varying lifetimes. Upon arrival of first responders, they quickly

assess casualties based on the same procedures for all MCIs, and categorize them. Hence, expected lifetime of different casualty groups is likely to be similar across different MCIs. On the other hand, service times can significantly vary, as a disaster can occur in a rural area with limited number of resources available (i.e., potentially longer travel distances), or in an urban area with a relatively higher availability of resources (i.e., potentially shorter travel times). Some studies in the literature such as Mills et al. (2013) analyze cases in which initial survival probability of a casualty type varies greatly (e.g., from 0.09 to 0.56). This results in a wide range of optimal policies, which can be quite different from one MCI to another. Assuming such unrealistic variation makes analysis of the triage problem rather arbitrary, and thus, we assume the main variation is regarding availability of resources that results is fluctuating service times. For simplicity, in the rest of our analysis, we use expected lifetimes for casualties of type 1 and 2 to be 200 and 300, respectively. In the next section, we study the service order problem with two casualty types, in order to compare and analyze the results from the introduced models.

3.4 Two Casualty Types

While there could be several categories of casualties in an MCI, in order to study the structure of the optimal policies and perform numerical analysis in a timely manner, we study cases with two casualty types throughout this paper. Although using two casualty types initially seems like a restriction, the same case happens in most scenarios in the practice. As mentioned earlier, START is the most common triage method and it has four casualty types, two of which are critical and compete for resources; Immediate (more critical) and delayed (less critical) casualties compete for resources, while minor casualties can wait longer and expectant has no chance of survival regardless of the resources available. Hence, we perform our analysis with two casualty types; type 1, more critical (e.g., immediate in START), and type 2, less critical (e.g., delayed in START). Type 2 casualties are still considered critical as they are in prompt need of medical treatment. As we examine this problem, we consider

two simple strategies, a strategy where all type 1 casualties are served first, followed by the type 2 casualties, which we denote as $S(1,2)$; this is the implied service order of the START triage system, and we also consider the opposite order, which we denote as $S(2,1)$.

Here, we illustrate and compare the results from our models. For Model 1, we have three state variables in the case with two casualty types: number of casualties of types 1 and 2 remaining to be served, and time. This means the results from Model 1 have three dimensions; the results indicate type of the casualty to be served when there are q_1 and q_2 casualties of types 1 and 2 are waiting to be served at time t. In case of Model 2, although there are three state variables the same as Model 1, the results can be illustrated only using q_1 and q_2, as we assume there is no information about casualties' death. Thus, all casualties are served and service time at each state can be determined using q_1 and q_2 under off-site death assumption as $s_1q_1 + s_2q_2$. In case of Model 3, it has a similar behavior to that of Model 1, as casualties' death is observed. In other words, results from Models 1 and 3 are stochastic, while for Model 2 it is deterministic. Figure 3.3 shows the partial decision tree for service to four casualties, two type 1 and two type 2, at time t'. At decision epoch t', there are two possible decisions, serving a type 1 or a type 2 casualty. Based on the decision, the three possible outcome are combined into two in the graph; service to the selected casualty type (i.e., on-site survival, that is,when casualty is alive, or when the casualty is not going to survive, but the it has not been observed), or no service (i.e., on-site death, that is, when it is observed that the casualty is not going to survive). The probability for the former outcome, as shown earlier in Expression 3.2), is $f_{it} + (1 - f_{it})(1 - O_{it})$ and for the latter is $(1 - f_{it})O_{it}$, assuming casualty type i is selected for service at epoch t. In Figure 3.3, only decisions and states that lead to having one casualty of each type are depicted. For Model 2, $O_{it} = 0$, that is, the "no service" branch has probability of 0 ($(1 - f_{it})O_{it} = 0$). Thus, at each decision epoch, we only have the upper arrow as the possible outcome, since no death is observe, and there is no stochasticity related to time of being at each state. In the example shown in Figure 3.3, if we remove all the outcomes regarding observing death, we can only be at state $(q_1, q_2) = (1, 1)$ at time $t = t' + s_1 + s_2$. For Model 3, $O_{it} = 1$, thus, probabilities

of traversing states change, but the overall structure of the decision tree remains exactly the same as Model 1. At each decision epoch, outcome of both decision are calculated using the displayed probabilities and collected rewards (not displayed in Figure 3.3). Then, all possible cases are analyzed until each branch reaches the state with no casualties remaining. Finally, the times with highest reward are picked for each state.

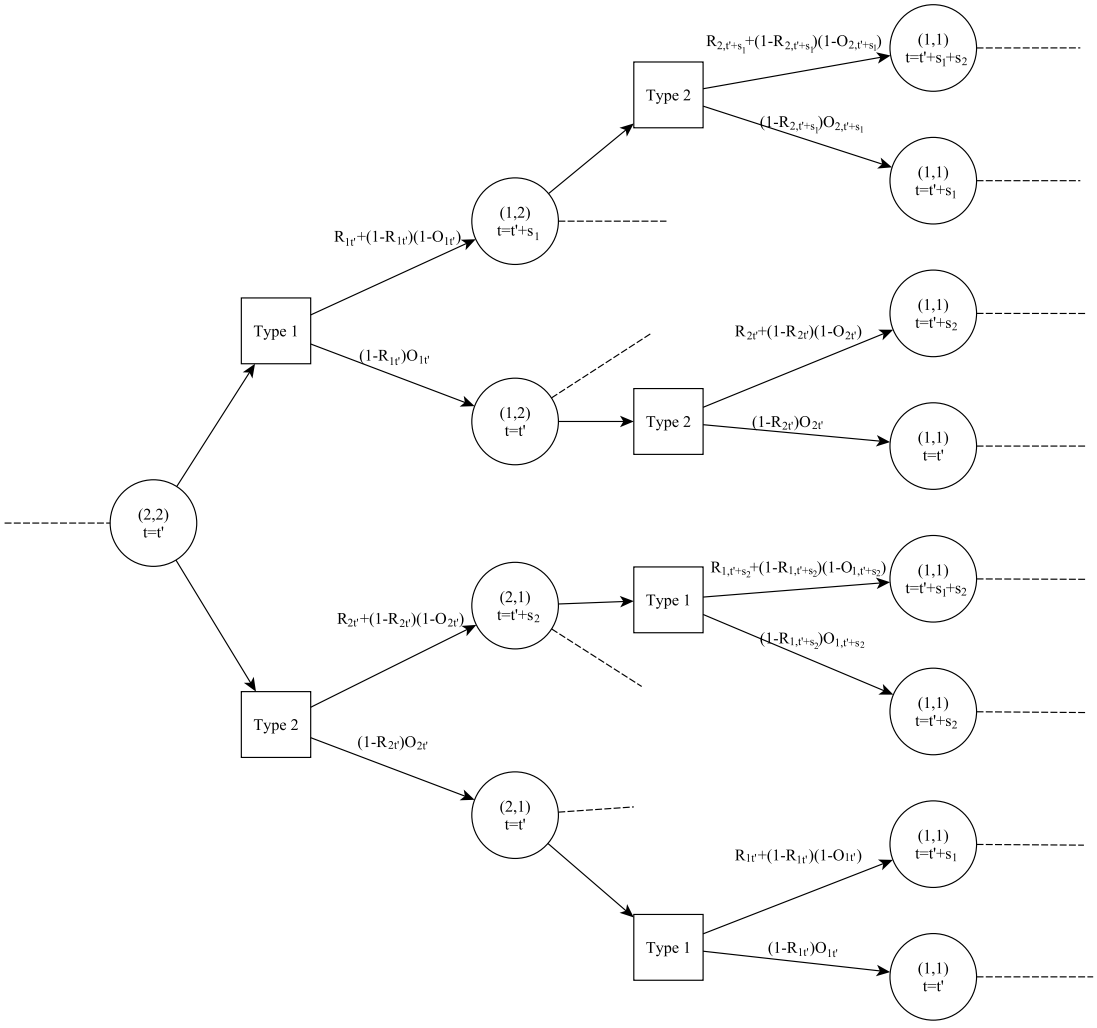


Figure 3.3: Partial decision tree of service to 4 casualties at time t' under Model 1

To further explain the stochasticity in the results, consider the case shown in Figure 3.3 with two casualties of each type. There are eight possible paths that lead to state $(q_1, q_2) = (1, 1)$ at four different time intervals. Assuming there are other times we can visit the (2,2) state,

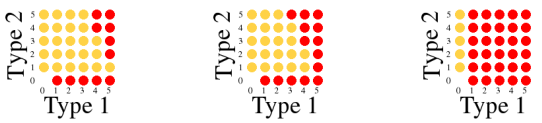
there will be more times we can be at (1,1). Therefore, due to death observability, optimal service type could be different at the same point based on the time, that is, the actual service graph is three dimensional for the case with two casualty types. We have a dimension for number of the remaining casualties of each type to be served and a time dimension. As mentioned earlier, another model in the literature that considers casualty's death (not observability) is developed by Jacobson et al. (2012). Their results are represented in two-dimensional format due to the memory-less property of their model (they use Exponential distribution for both lifetimes and service times). Our deterministic service times increases the complexity of the results, but it provides a better estimation and allows for studying death observability. In this section, for demonstrating the results from Models 1 and 3, we show the *maximum likelihood* projection of service orders, but later in our numerical analysis we show the breakdown on service order based on our simulation analysis.

In the results, we observe that at most we have one service change for different times at the same point. In other words, either service order remain the same during different possible time intervals for a casualty state, or it changes at most once from wither type 1 to type 2 or the other way. Table 3.1 shows the optimal service for a problem with initially 5 casualties of each type with expected lifetimes of 200 and 300, service times of 20 and 15, and rewards of 1. When there are 2 casualties of each type are remaining, possible time periods are listed in the top row of Table 3.1, and the corresponding value in the bottom row shows the casualty type that should be served. This instance shows how have at most one service type switch at one casualty state. Thus, the maximum likelihood is an appropriate approach for representing the three dimensional results. The maximum likelihood graph shows the optimal decision that is most likely to occur among different time intervals at a certain state of the remaining casualties. In order to generate the maximum likelihood service graph, we calculate the probability of being at each time period for a certain casualty state (e.g., (q_1, q_2)), then, we calculate sum of the probabilities for times with the same service order. The service order with the higher probability determines the maximum likelihood result for that casualty state. We use this approach to demonstrate the results from Models 1 and 3 with two dimensions;

Table 3.1: Service type for a case with (2,2) casualties remaining from initial (5,5) casualties abd their associated probability with $(s_1, s_2) = (20, 15)$

	Model 1		Model 2		Model 3	
Time	Service	Probabilities	Service	Probabilities	Service	Probabilities
15	1	0.00	-	-	1	0.00
20	1	0.00	-	-	1	0.00
30	1	0.00	-	-	1	0.00
35	1	0.00	-	-	1	0.00
40	1	0.00	-	-	1	0.00
45	1	0.00	-	-	1	0.00
50	1	0.00	-	-	1	0.00
55	1	0.00	-	-	1	0.00
60	1	0.00	-	-	1	0.00
65	1	0.00	-	-	1	0.01
70	1	0.00	-	-	1	0.02
75	2	0.00	-	-	1	0.03
85	2	0.02	-	-	1	0.09
90	2	0.04	-	-	1	0.12
105	2	0.35	2	1	1	0.35

number of type 1 and type 2 casualties waiting for service. Otherwise, we need a three-dimensional representation with a time axis. For Model 2, since we serve all casualties, each casualty state is associated with exactly one time interval $((n_1 - q_1)s_1 + (n_2 - q_2)s_2)$. Thus, the results from Model 2 can be represented in a two-dimensional format without any transformation. In fact, in Proposition 6 we show that we can obtain the results for Model 2 from the results of Model 1, by picking the optimal service type for the largest time interval at each casualty state, as we serve all casualties regardless of their death before receiving service and there is no possibility of being at any smaller time interval for a casualty state.



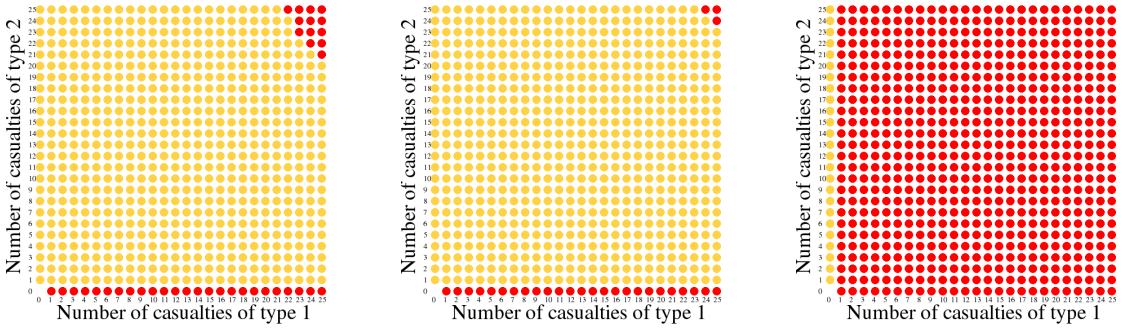
Model 1 Model 2 Model 3

Figure 3.4: Results from Models 1, 2, and 3 with $n_1 = n_2 = 5$, equal rewards $((\alpha_1, \alpha_2) = (1.00, 1.00))$, $(s_1, s_2) = (20, 15)$, $(l_1, l_2) = (200, 300)$, and $(d_1, d_2) = (200, 200)$ for Model 1.

Proposition 6 *We can obtain the results for Model 2 from the results of Model 1, by picking the optimal service type for the largest time interval at each casualty state.*

Proof. See Appendix.

Figure 3.5 illustrates the maximum likelihood service graph form Models 1, 2, and 3 for a scenario with 25 casualties of each type, $(l_1, l_2) = (200, 300)$, $(s_1, s_2) = (20, 18)$, and $(\alpha_1, \alpha_2) = (1.00, 1.00)$. For Model 1, we use $(d_1, d_2) = (200, 200)$. In order to interpret the *service graph*, we need to define the *service path*. A path for a case with n_i casualties of each type for $i \in \{1, 2\}$, starts at the top right corner of the graph. Then, the service path continues through the graph until all casualties are served and ends at bottom left corner. At each point, color of the circle determines the type of the casualty to be served (transported). A dark (red) circle in the graph indicates serving a casualty of type 1 (e.g., immediate in START) and light (yellow) circle indicates serving a casualty of type 2 (e.g., delayed in START). The state of the system changes based on the service type and observed deaths. After determining the state of the system, the point on the graph for the current state indicates the service type. The service continues until no casualty is left.

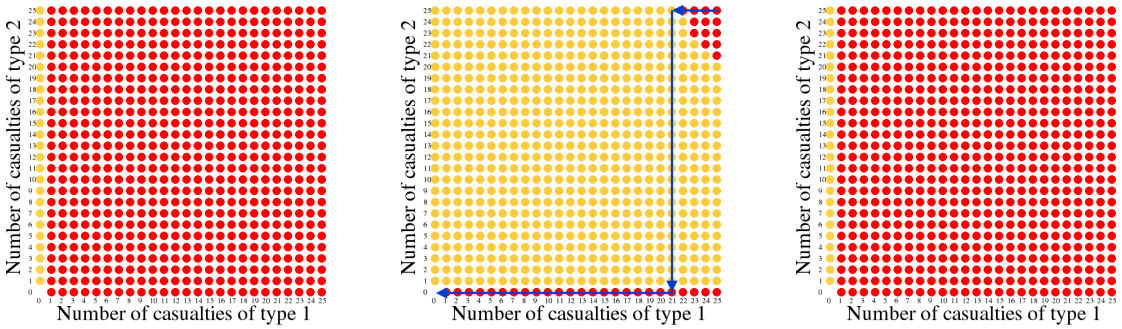


Model 1 Model 2 Model 3

Figure 3.5: Results from Models 1, 2, and 3 with $n_1 = n_2 = 25$, equal rewards $((\alpha_1, \alpha_2) = (1.00, 1.00))$, $(s_1, s_2) = (20, 18)$, $(l_1, l_2) = (200, 300)$, and $(d_1, d_2) = (200, 200)$ for Model 1.

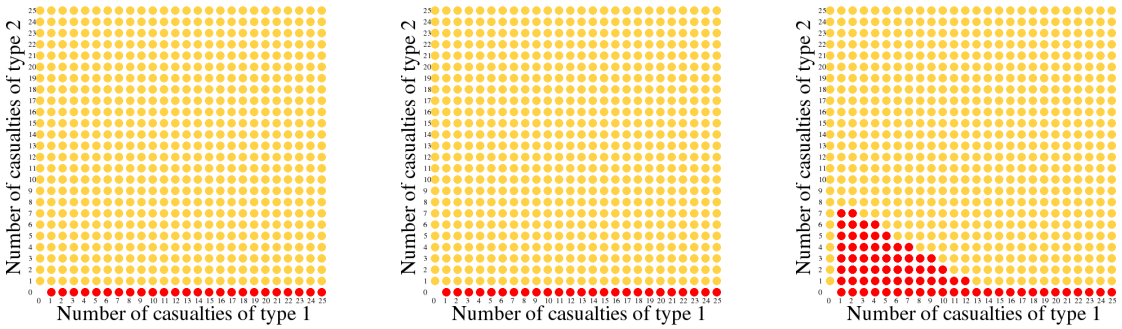
Figure 3.4 highlights the difference in the results from our models under the same scenario. Both Models 1 and 2 generate similar results under this scenario, serving a few type 1

casualties initially, all type 2 casualties, and then remaining type 1 casualties. On the other hand, Model 3 generates a $S(1,2)$ policy similar to START. Under Model 2, $d_i \to \infty$, that is, we expect to receive no information regarding casualties' death (i.e., receive information at infinity). As we decrease the d_i values, the results shift towards serving more type 1 casualties, where it becomes $S(1,2)$ under Model 3 (with $d_i = 0$). These strategies highlight the importance for a flexible model, to generate dynamic policies for serving casualties in various scenarios.



Model 1 Model 2 Model 3

Figure 3.6: Results from Models 1, 2, and 3 with $n_1 = n_2 = 25$, equal rewards $((\alpha_1, \alpha_2) = (1.00, 1.00))$, $(s_1, s_2) = (20, 20)$, $(l_1, l_2) = (200, 300)$, and $(d_1, d_2) = (200, 200)$ for Model 1.



Model 1 Model 2 Model 3

Figure 3.7: Results from Models 1, 2, and 3 with $n_1 = n_2 = 25$, equal rewards $((\alpha_1, \alpha_2) = (1.00, 1.00))$, $(s_1, s_2) = (20, 15)$, $(l_1, l_2) = (200, 300)$, and $(d_1, d_2) = (200, 200)$ for Model 1.

Figures 3.6 and 3.7 illustrate the results from all three models under different scenarios with $(s_1, s_2) = (20, 20)$ and $(s_1, s_2) = (20, 15)$, respectively. In Figure 3.6 Models 1 and 3 both generate $S(1,2)$, while in Figure 3.7 Models 1 and 2 both generate $S(2,1)$. Across

different scenarios, as we increase criticality (i.e., decrease service time gap by increasing s_2), results in all models shift from $S(2,1)$ towards $S(1,2)$. This indicates that in highly resource-constrained scenarios, $S(1,2)$-like policies perform better, and the opposite in less resource-constrained ones. Under the off-site death assumption, the dark (blue) arrows in Figure 3.6 show the optimal service path (i.e., the optimal plan). We can think of the optimal solution as a vector of 1's (or dark circles) and 2's (light circles) along the service path, representing the order that casualties of each type should be served.

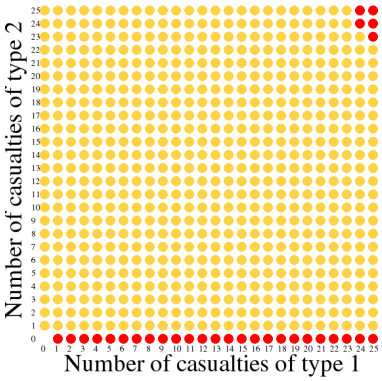
We chose expected values for d_1 and d_2 in above analysis based on the empirical data presented in Frykberg and Tepas III (1988). As mentioned earlier, Frykberg and Tepas III (1988) found a very small number of deaths among surviving casualties following an MCI prior to receiving service. Hence, use of larger d_i values is more justified. To study how changing d_i values affect the results, in Figure 3.8 we illustrate the optimal service graph for the same scenarios as Figure 3.4 with $(d_1, d_2) = (100, 100)$ and $(d_1, d_2) = (1000, 1000)$. It can be observed that decreasing d_i shifts the results towards $S(1,2)$. In the first case, expected time of observing death is 1000, compared to Figure 3.4, which is 200, thus, fewer type 1 casualties are served initially. For the second case, expected time of observing death is 100, and the optimal policy is similar to that of $S(1,2)$, except a few type 2 casualties shortly after starting service. Next, we look into analytical properties of the proposed models to gain some insight into the optimal policies.

3.5 Analytical Results

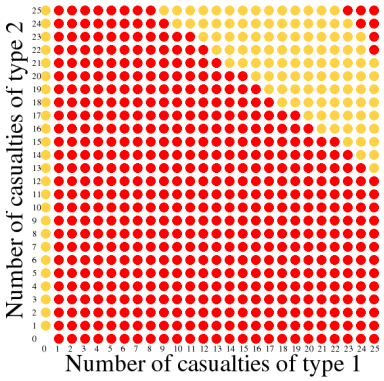
In this section we examine the solution properties for our proposed Models.

Proposition 7 *Assuming a finite number of casualty types, the worst-case complexity of Model 1 is polynomial.*

Proof. See Appendix.



$(d_1, d_2) = (1000, 1000)$



$(d_1, d_2) = (100, 100)$

Figure 3.8: Results from Model 1 with $n_1 = n_2 = 25$, equal rewards ($(\alpha_1, \alpha_2) = (1.00, 1.00)$), $(s_1, s_2) = (20, 18)$, and $(l_1, l_2) = (200, 300)$.

Proposition 7 indicates that Model 1 can be solved in a timely manner, and size the state space does not grow exponentially in relation to problem parameters. Although the growth is not exponential, we found that the Model 1 cannot be solved for practical scenarios. For instance, it takes more than an hour to solve a case with five casualty types and 10 casualties of each type. Models 1 formulates the service order problem considering casualty death both on-site, prior to receiving service, and off-site, after receiving service.

Mills et al. (2013) show that in the optimal solution to the service order problem under off-site death assumption with equal service times, once service to type 2 casualties start, it is not interrupted unless all type 2 casualties are served. As stated earlier, Kamali et al. (2016) proves that the problem under the mentioned setting with equal service times becomes a knapsack problem that can be solved to optimality using a simple greedy algorithm. In addition, Kamali et al. (2016) proves the mentioned property of the optimal solution under off-site death assumption also holds when service times are not equal. Here, we extend the finding by showing that the solution structure holds even with unequal reward values for each casualty type.

Proposition 8 *There exists an optimal solution to Model 2 under Assumption 3, where*

service of a type 2 casualty once, started, is not interrupted.

PROOF. See Appendix.

This proposition implies that, in an optimal solution under the off-site death assumption, service to type 2 casualties is not interrupted once started. Thus, in an optimal solution the service order is in one of the following forms:

- All type 1 casualties are served first, and then the type 2 casualties ($S(1,2)$).
- A portion of the type 1 casualties are served first, and then all type 2 casualties are served, followed by any remaining type 1 casualties.
- All type 2 casualties are served first, and then all remaining type 1 casualties ($S(2,1)$) are served.

This can be thought of as three intervals for serving casualties. The first and third intervals are for serving type 1 casualties and each can serve from 0 to n_1 casualties, but their total should always add to n_1. Second interval is for serving less-critical casualties and always serves all n_2 casualties.

Next, we study Model 3, which produces the optimal policy under the *on-site death* assumption. In Proposition 9, we show that when the service times and rewards are equal, the optimal policy is equivalent to $S(1,2)$.

Proposition 9 *The optimal solution to Model 3, when service times and rewards for both casualty types are equal (i.e., $s_1 = s_2$ and $\alpha_1 = \alpha_2$) is to serve the type 1 casualties first, and then type 2 casualties (i.e., $S(1,2)$).*

PROOF. See Appendix.

Proposition 9 shows that Model 3 under the setting proposed by Mills et al. (2013) results into the static $S(1,2)$ policy. Due to longer stabilization time required for more critical casualties, it is unlikely for service times to be equal in practice.

Proposition 10 *The model proposed by Jacobson et al. (2012) cannot have a time-dependent reward.*

Proof. See Appendix.

Proposition 10 shows that the model developed by Jacobson et al. (2012) under on-site death assumption cannot have a time-dependent reward. Our models can simply be modified to have a time-dependent reward (α_{it}), if needed. One case such property is needed, is when using quality-adjusted life-year (QALY) as the reward. Depending on the service time, the QALY score varies for each casualty type, which can be incorporated in our models. Model 1 (and Models 2 and 3) can be used with any distribution, but since it uses deterministic service times, it does not show memory-less property. Next we explore the use of exponential lifetimes, as proposed by Jacobson et al. (2012). Assuming we have two casualty types with expected lifetimes l_1 and l_2, without loss of generality we assume type 1 is the more critical casualty type. This means $l_1 \leq l_2$. Thus, the survival probability functions are:

$$f_{it} = e^{-\frac{t}{l_i}}, \qquad i = 1, 2$$

When survival probabilities follow exponential distribution, Assumption 1 holds. Since we have $l_1 \leq l_2$ and $f_{1t} \geq f_{2t}$, we can find the difference of two survival probabilities as follows:

$$f_{2t} - f_{1t} \;=\; e^{-\frac{t}{l_2}} - e^{-\frac{t}{l_1}} \tag{3.8}$$

We need to show expression (3.8) has a unique maximum.

$$\frac{d(f_{2t} - f_{1t})}{dt} = -\frac{1}{l_2} e^{-\frac{t}{l_2}} + \frac{1}{l_1} e^{-\frac{t}{l_1}}$$

Based on definition of types, we know that $l_2 \geq l_1$ and hence, $\frac{d(f_{2t}-f_{1t})}{dt} > 0$ at $t = 0$. On the other hand, $\lim_{t\to-\infty} \frac{d(f_{2t}-f_{1t})}{dt} = 0^+$ and $\lim_{t\to+\infty} \frac{d(f_{2t}-f_{1t})}{dt} = 0^-$. Therefore, $f_{2t} - f_{1t}$ has one extremum, which is a maximum.

Proposition 11 *If the survival probabilities follow the Exponential distribution having l_1 and l_2 as the means for the two casualty types, the optimal switching point, $n_1'^*$, can be obtained from equation (3.9) under the off-site death assumption.*

$$n_1'^* = \frac{l_1 l_2}{s_1(l_1 - l_2)} \ln \left(\frac{\alpha_2 l_1}{\alpha_1 l_2} \frac{\frac{1-e^{-\frac{n_2 s_2}{l_2}}}{1-e^{-\frac{s_2}{l_2}}}}{\frac{1-e^{-\frac{n_2 s_2}{l_1}}}{1-e^{-\frac{s_1}{l_1}}}} \right). \tag{3.9}$$

Proof. See Appendix.

In expression (3.9), if $n_1'^* \geq n_1$, it implies serving all type 1 casualties first, and then all type 2 casualties, as no delays in service are allowed. If $0 < n_1'^* < n_1$, then, $n_1'^*$ type 1 casualties should be served, followed by all type 2, and finally rest of type 1. If $n_1'^* \leq 0$, then all type 2 casualties should be served first, followed by all type 1. In terms of number of casualties, expression (3.9) is only dependent on n_2, which supports our earlier finding in the Proposition 8. In other words, we only need number of type 2 casualties to find the time to start their service under off-site death assumption. However, in practice we need number of type 1 casualties, as if their service finishes before $n_1'^*$, we need to start service for type 2 casualties earlier. Proposition 11 shows if the setting proposed by (Jacobson et al., 2012) is solved under off-site death assumption, the optimal policy can be found using a simple expression. In our numerical analysis, we compare the performance of results found using

this expression as a heuristic to that of other models.

3.6 Numerical Analysis

In this section, we compare the solutions from the above described models and heuristics numerically through simulation. In the simulation analysis we assess the performance of the various solutions including the ones under the on-site death and off-site death assumptions. For the Model 1, when a deceased casualty (not served before his or her lifetime, based on the underlying probability distribution used) of type i, $i \in \{1, 2\}$, is to be served, there is a $1-O_{it}$ probability that the casualty's death is not observed, which implies that resources are used, but no reward is collected, and a O_{it} probability that the death is observed, and thus is not served and no reward is collected. For lifetimes, we consider exponential distribution as well as the distribution Sacco et al. (2005) derived from the data. The simulations, as per the problem description, has two casualty types, type 1 (more critical) and type 2 (less critical) and one server (i.e., one emergency vehicle) serving both casualty types. Each scenario is simulated 500 times and we report number of time each method generates the policy having the highest objective function value (reported as *Times Best* in the various tables), as well as the the average number of the type 1, type 2, and total casualties that survive. The breakdown of survivors by type helps us better understand the structure of the policies used.

First, we compare the results from $S(1, 2)$ and $S(2, 1)$ strategies, Models 1, 2, 3, and Model 3 with observed deaths. We also include Model 3 with observed deaths to make Model 3 more similar to the reality. Model 3 as presented earlier has two terms; service if the casualty is not dead at the time of the service, and removal from the system if the casualty is dead at the time of the service. For Model 3 with observed deaths we replace the second term with the probability of observed death ($(1 - f_{it})O_{it}$)). We do not include Expression (3.9) in the first analysis, as it generates the same results as Model 2. Results from the heuristic methods and

Expression (3.9) are compared later. We use five scenarios, in which we vary the expected service times of type 2 casualties from shorter to longer, keeping lifetimes and expected type 1 service time the same. Expected lifetimes are 200 and 300 for type 1 and type 2 casualties, respectively. Other related studies in the literature vary the expected lifetime of casualties in their numerical analysis and keep the expected service times constant as explained earlier. There are 25 casualties of each type, and we generate lifetimes randomly from the expected values. We generate service times for each casualty randomly from log-normal distribution with with the given mean and standard deviation, as suggested by Ingolfsson et al. (2008). Based on the data provided by Frykberg (2002), we use 200 as the expected time for observing death, d_i, in all scenarios.

Table 3.2 shows as expected Model 1 is not outperformed in any of the scenarios. In the first scenario (least critical) $S(2,1)$ is the optimal policy, and on the last (most critical) $S(1,2)$ is the optimal. Looking at the breakdown of the casualties survived, we can observe that the results from Model 2 tends towards $S(2,1)$ compared to Model 1, while for Model 3 results tends towards $S(1,2)$. When we change Model 3 to account for observed deaths, its results shift towards $S(2,1)$ similar to that of Model 1. As shown in Proposition 9, when service times are equal (last scenario), Model 3 results in $S(1,2)$. In all scenarios $S(1,2)$ results in largest total unobserved deaths. Although Model 1 generates the largest total survivors in all scenarios, in some scenarios it does not generate smallest total unobserved deaths. Model 1 does not generate the best results in every simulation run due to the stochasticity in the population lifetimes and service times. To further show the effect of stochasticity, for the case with $s_1 = 20, s_2 = 18$ in the simulation (optimal solution in Figure 3.4 for Model 1), we have counted number of the times each state is visited for each type of service. Table 3.3 shows number of the time type 1 casualties are served at each state out of 500 simulation runs, and Table 3.4 shows the same for type 2 casualties. These tables highlight the importance of the service time in determining the optimal service type, as well. For instance, in state $(q_1, q_2) = (22, 25)$, 108 times a type 1 casualty is served and 24 times a type 2, as service type depends on the epoch. Moreover, these tables show the expected path that is traveled

Table 3.2: Simulation of the results from $S(1,2)$, $S(2,1)$, Models 1, 2, and 3 with $n_1 = n_2 = 25$, $(l_1, l_2) = (200, 300)$, $\alpha_1 = \alpha_2 = 1$, and $d_1 = d_2 = 200$

		Survived			Unobserved Death		
Method	Best	Type 1	Type 2	Total	Type 1	Type 2	Total
			$s_1 = 20, s_2 = 16$				
$S(1,2)$	152	41.75	25.26	33.51	23.47	11.10	17.28
$S(2,1)$	291	14.30	57.31	35.81	14.67	18.47	16.57
Model 1	291	14.30	57.31	35.81	14.67	18.47	16.57
Model 2	291	14.30	57.31	35.81	14.67	18.47	16.57
Model 3	180	26.30	42.91	34.61	21.91	12.30	17.11
Model 3 wOD	291	14.30	57.31	35.81	14.67	18.47	16.57
			$s_1 = 20, s_2 = 17$				
$S(1,2)$	144	41.75	24.90	33.32	23.47	11.02	17.24
$S(2,1)$	197	13.62	55.85	34.73	14.15	18.47	16.31
Model 1	215	16.80	52.97	34.88	13.67	18.97	16.32
Model 2	197	13.62	55.85	34.73	14.15	18.47	16.31
Model 3	153	34.53	32.47	33.50	23.86	10.62	17.24
Model 3 wOD	197	13.62	55.85	34.73	14.15	18.47	16.31
			$s_1 = 20, s_2 = 18$				
$S(1,2)$	215	41.75	24.39	33.07	23.47	11.18	17.33
$S(2,1)$	168	13.04	54.33	33.68	13.63	18.53	16.08
Model 1	196	20.87	46.95	33.91	13.64	18.37	16.00
Model 2	172	17.54	50.26	33.90	13.26	18.58	15.92
Model 3	215	41.75	24.39	33.07	23.47	11.18	17.33
Model 3 wOD	168	13.04	54.33	33.68	13.63	18.53	16.08
			$s_1 = 20, s_2 = 19$				
$S(1,2)$	230	41.75	23.98	32.87	23.47	11.10	17.28
$S(2,1)$	187	12.41	53.01	32.71	13.08	18.57	15.82
Model 1	245	39.82	26.28	33.05	22.03	11.82	16.93
Model 2	199	18.20	47.78	32.99	13.10	18.50	15.80
Model 3	230	41.75	23.98	32.87	23.47	11.10	17.28
Model 3 wOD	187	12.41	53.01	32.71	13.08	18.57	15.82
			$s_1 = 20, s_2 = 20$				
$S(1,2)$	264	41.75	23.64	32.70	23.47	11.06	17.26
$S(2,1)$	136	11.82	51.93	31.88	12.84	18.27	15.56
Model 1	264	41.75	23.64	32.70	23.47	11.06	17.26
Model 2	176	18.94	45.71	32.32	12.66	18.31	15.48
Model 3	264	41.75	23.64	32.70	23.47	11.06	17.26
Model 3 wOD	159	14.73	49.55	32.14	12.56	18.39	15.48

through the service graph, which is the results of removal of casualties with observable death from the system.

Table 3.3: Number of times each state is visited for serving type 1 casualties in 500 runs of the simulation from Table 3.2 with $s_1 = 20, s_2 = 18$

Type	Type 1																								
2	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
25																			2	1	8	30	108	169	310	500
24																				1	6	19	49	76	51	
23																	1					10	14	15	3	
22																1		1			1	2	1	1		
21																				2						
20																					1					
19																						1				
18																1										
17																										
16																										
15																										
14																										
13																										
12																1										
11																										
10																										
9					1				1																	
8		1				1		1																		
7			1	1				1					1													
6					2		1	1	1																	
5		2	2	1	3	3	4	4	1	2																
4		1	1	2	5	2	4	2	1	2																
3		2	4	5	5	4	3	2	2	2																
2		4	3	8	5	4	6	9	2	2	4		2													
1		4	5	7	3	11	12	6	1	3	2															
0		214	198	172	155	135	121	89	71	48	23	17	3	3	1	1										

3.6.1 Heuristics

Our developed models are DPs and they become less tractable as problem size grows. Here, we analyze the performance of Expression (3.9), which generates the optimal solution under off-site death assumption, as a heuristic for the general case. This expression assumes exponential lifetimes for both casualty types, and gives the optimal service order by just solving an expression. Jacobson et al. (2012) proposes four different heuristic approaches to solve this problem under on-site death assumption (i.e., $O_{it} = 1$). We describe two of the heuristics, the *Myopic policy* and the $\alpha r\mu$*-rule*, both of which are state-independent, that is, they are not dependent on number of the casualties in each category. The other two heuristics are more complicated to implement and we can find no justification to use them instead

Table 3.4: Number of times each state is visited for serving type 2 casualties in 500 runs of the simulation from Table 3.2 with $s_1 = 20, s_2 = 18$

Type	Type 1																									
2	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
25																			4	17	40	44	24			
24																	6	8	16	37	62	69	32			
23															3	4	5	10	28	48	69	51	21			
22														2	1	2	10	23	29	51	54	46	17			
21													1		2	12	12	26	46	50	50	23	8			
20													1	1	6	11	21	35	59	59	43	11	2			
19												2	2	5	10	21	20	37	49	43	22	7	4			
18											1		2	5	14	28	34	43	56	25	17	8	2			
17										2		3	5	10	20	35	38	51	35	25	6	2	1			
16											1	3	14	14	32	24	34	31	32	15	6	3				
15										2	3	6	14	23	32	45	48	37	24	12	6	2				
14									2		7	17	21	19	28	31	34	29	18	4	3					
13								1	3	6	7	14	22	31	40	42	24	26	13	8	1	1				
12							2	1	7	7	8	23	30	33	36	31	39	14	7	4	1					
11							3	3	9	8	22	32	30	34	44	30	23	17	5	2						
10							4	2	9	10	25	35	31	35	33	23	15	11	5							
9						3	3	10	10	19	22	39	38	36	23	23	13	6	2							
8					2	1	2	10	16	21	38	43	36	27	29	9	13	4	1							
7	1	1		1		9	6	16	20	28	37	33	36	28	13	6	9	1	1							
6	1	1		2	3	10	9	12	25	41	48	42	34	24	13	11	4	2								
5	2	2		3	9	13	8	18	27	41	52	44	26	18	11	8		3	1							
4	2	2		6	6	15	15	31	39	54	41	34	16	22	6	6		1								
3	6	6	4	5	8	16	20	36	56	47	38	30	22	10	6	3	2									
2	14	3	12	11	12	21	23	47	50	36	36	28	18	9	2	2	1									
1	21	11	11	12	26	33	28	36	54	45	21	21	7	4	3	1	1									
0																										

of Model 3, which solves the problem under on-site death assumption to optimality. The Myopic policy gives priority to the casualty type $i \in \{1, 2\}$ that maximizes $\frac{\alpha_i}{l_i}/(\frac{1}{s_{(3-i)}} + \frac{1}{l_i})$, while the $\alpha r\mu$-rule gives priority to the casualty type $i \in \{1, 2\}$ that maximizes $\frac{\alpha_i}{l_i s_i}$. Here, we compare the results from this expression to that of heuristics introduced by Jacobson et al. (2012), $S(1, 2)$, and $S(2, 1)$, all of which are simple and easy to implement policies.

Table 3.5 shows that Expression (3.9) performs equal to or better than other methods under different scenarios. This advantage is achieved by the ability of Expression (3.9) to generate mixed policies, rather than static ones similar to $S(1, 2)$ and $S(2, 1)$. performance of $S(2, 1)$ degrades as scenarios become more resource-constrained, while performance of $S(1, 2)$ increases, and it finally outperforms $S(2, 1)$ in the final scenario. Both heuristics from Ja-

Table 3.5: Simulation of the results from $S(1,2)$, $S(2,1)$, Expression (3.9), and heuristics from Jacobson et al. (2012) with $n_1 = n_2 = 25$, $(l_1, l_2) = (200, 300)$, and $\alpha_1 = \alpha_2 = 1$

Method	Best	Type 1	Type 2	Total
		$s_1 = 20, s_2 = 13$		
$S(1,2)$	99	41.75	26.49	34.12
$S(2,1)$	437	16.61	62.20	39.40
Expression (3.9)	437	16.61	62.20	39.40
$\alpha r \mu$	437	16.61	62.20	39.40
Myopic	437	16.61	62.20	39.40
		$s_1 = 20, s_2 = 17$		
$S(1,2)$	227	41.75	24.90	33.32
$S(2,1)$	332	13.62	55.85	34.73
Expression (3.9)	332	13.62	55.85	34.73
$\alpha r \mu$	227	41.75	24.90	33.32
Myopic	227	41.75	24.90	33.32
		$s_1 = 20, s_2 = 18$		
$S(1,2)$	230	41.75	24.39	33.07
$S(2,1)$	192	13.04	54.33	33.68
Expression (3.9)	230	19.15	48.78	33.96
$\alpha r \mu$	230	41.75	24.39	33.07
Myopic	230	41.75	24.39	33.07
		$s_1 = 20, s_2 = 19$		
$S(1,2)$	252	41.75	23.98	32.87
$S(2,1)$	195	12.41	53.01	32.71
Expression (3.9)	206	21.66	44.57	33.11
$\alpha r \mu$	252	41.75	23.98	32.87
Myopic	252	41.75	23.98	32.87
		$s_1 = 20, s_2 = 20$		
$S(1,2)$	269	41.75	23.64	32.70
$S(2,1)$	181	11.82	51.93	31.88
Expression (3.9)	200	24.20	40.74	32.47
$\alpha r \mu$	269	41.75	23.64	32.70
Myopic	269	41.75	23.64	32.70

cobson et al. (2012) generate $S(2,1)$ in the first three scenarios (less resource-constrained) and $S(2,1)$ in the last two scenarios (more resource-constrained). This table highlights the importance of having heuristics that are able to generate a wide variety of policies under different scenarios.

As mentioned earlier, Mills et al. (2013) propose two heuristics to address the problem: QS-ReSTART and QD-ReSTART. QS-ReSTART is simple heuristic generating either $S(1,2)$ (all type 1 casualties first, followed by all type 2 casualties) or $S(2,1)$ (all type 2 casualties first, followed by all type 1 casualties), while QD-ReSTART is more advanced and can generate more complex results as well. Kamali et al. (2016) demonstrate that their proposed model under off-site death assumption solves the fluid formulation developed by Mills et al. (2013) to optimality, and outperforms the mentioned heuristics in a numerical analysis. Hence, as a results of Proposition 3, Model 2 also provides the optimal solution to the fluid formulation developed by Mills et al. (2013). In fact, the problem under Mills et al. (2013) assumptions with two casualty types and *equal service times* is equivalent to a simple knapsack problem (see Kamali et al., 2016, for details). The equivalent knapsack problem can be solved to optimality using a straight-forward greedy search, thus, there is no need to solve either the fluid formulation or the proposed heuristics. Thus, we provide no comparison of the results with Mills et al. (2013) heuristics as our Model 2 solve the same problem to optimality.

3.6.2 Sensitivity Analysis

In this section we analyze the sensitivity of the results to the expected death observation time, d_i. Our previous results are using $d_i = 200$. Here, we vary d_i values from 100 to 500, studying cases with equal and unequal values. In all cases we keep d_1 at most equal to d_2, as we expect to observe type 1 casualties' death earlier due to their worse initial condition. All the simulation settings are similar to that of Table 3.2, except we keep the service times $(s_1, s_2) = (20, 18)$.

Table 3.6 shows while results are relatively sensitive to d_i values, Model 1 generates the best results under all scenarios. As we increase d_i values, total number of survivors decrease due to increase in the resources time lost for serving expectant (unobserved dead) casualties. When $d_1 < d_2$, $S(1,2)$ tends to perform better, as expectant type 1 casualties are removed from the system at a faster pace. In general, as we increase d_i values across the scenarios,

Table 3.6: Sensitivity analysis of the results from $S(1,2)$, $S(2,1)$, and Model 1 to d_i values with $n_1 = n_2 = 25$, $(l_1, l_2) = (200, 300)$, $\alpha_1 = \alpha_2 = 1$, and $(s_1, s_2) = (20, 18)$

	Survived			
Method	Best	Type 1	Type 2	Total
	$d_1 = 100, d_2 = 100$			
$S(1,2)$	220	44.06	28.30	36.18
$S(2,1)$	175	16.21	56.14	36.17
Model 1	224	41.48	31.20	36.34
	$d_1 = 100, d_2 = 200$			
$S(1,2)$	169	44.06	26.89	35.48
$S(2,1)$	112	14.53	54.33	34.43
Model 1	180	36.06	35.46	35.76
	$d_1 = 200, s_2 = 200$			
$S(1,2)$	215	41.75	24.39	33.07
$S(2,1)$	168	13.04	54.33	33.68
Model 1	196	20.87	46.95	33.91
	$d_1 = 200, d_2 = 500$			
$S(1,2)$	188	41.75	21.57	31.66
$S(2,1)$	115	10.67	52.25	31.46
Model 1	188	23.14	41.54	32.34
	$d_1 = 500, d_2 = 500$			
$S(1,2)$	198	39.18	18.58	28.88
$S(2,1)$	205	8.40	52.25	30.32
Model 1	244	14.75	46.66	30.70

results transition from $S(1,2)$-like to $S(2,1)$-like, as observed earlier in the graphical representation of the optimal solution. Model 1, as expected, outperforms other strategies, but the relationship between d_i values is a determining factor in the structure of the optimal solution.

3.6.3 Multiple Casualty Types

While in majority of disasters there are two casualty types competing for resources, there are cases that some groups of casualties do not fall under any of these two categories. An example of this is a case with pediatric casualties, as they have different characteristics and require different initial care for stabilization. There are other cases such as burn victims or

patients with contagious diseases that might require different initial care. In order to study how having more than two casualty types affect the results, next we study cases with three and four casualty types. Table 3.7 shows the results for a case with four casualty types with $n_i = 10$, $s_i = 20$ and $d_i = 200$ for $i = 1, \ldots, 4$. Expected lifetimes are 200 for casualties of type 1 and 2, and 300 for types 3 and 4, and $\alpha_1 = \alpha_3 = 1$ and $\alpha_2 = \alpha_4 = 0.8$.

Table 3.7: Simulation of the results from $S(1,2,3,4)$, $S(4,3,2,1)$, Models 1, 2, and 3 with 10 casualties of each type, $l_1 = l_2 = 200$, $l_3 = l_4 = 300$, $\alpha_1 = \alpha_3 = 1$, $\alpha_2 = \alpha_4 = 0.8$, and $d_1 = d_2 = d_3 = d_4 = 200$

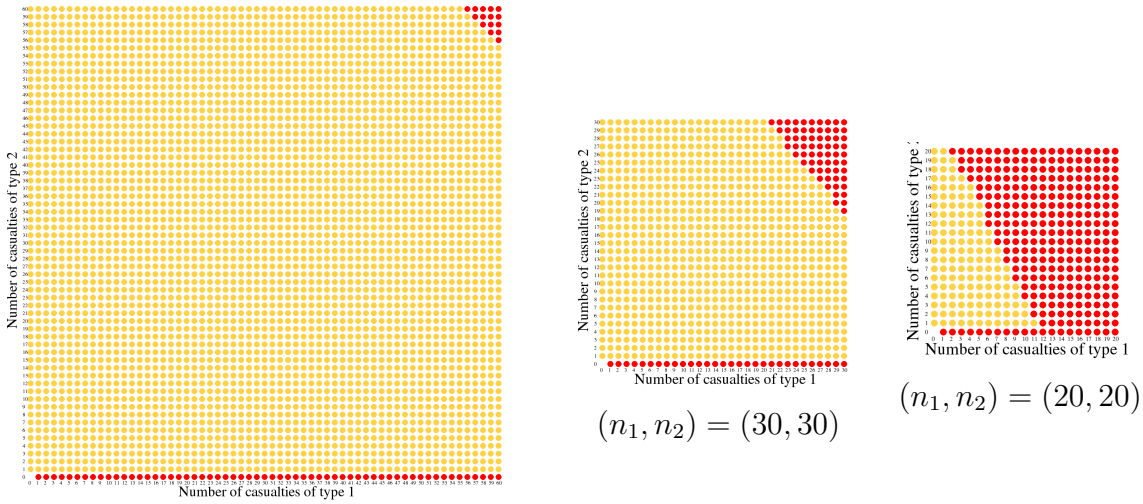
Method	Best	Type 1	Type 2	Type 3	Type 4	Total
	$s_1 = s_2 = s_3 = s_4 = 20$					
$S(1,2,3,4)$	168	6.69	2.82	2.38	1.65	13.55
$S(4,3,2,1)$	162	0.61	1.28	3.99	7.59	13.47
Model 1	195	5.52	0.71	4.80	2.84	13.88
Model 2	178	6.45	0.67	4.18	2.55	13.84
Model 3	172	6.69	2.74	2.48	1.65	13.56

Table 3.7 shows as expected Model 1 outperforms other approaches, while Models 2 and 3 both outperform $S(1,2)$ and $S(2,1)$. Model 1, and to a lesser extent Models 2 and 3, give priority to type 3 with longer expected lifetimes and higher reward. Same as before, results from Model 3 is closer to $S(1,2,3,4)$ than other models, but it is not exactly the same as rewards are not equal (see Proposition 9).

3.7 Multiple Servers

Our models are developed with the assumption of one server, serving casualties in a non-preemptive manner. This assumption may not be realistic, but it allows us to gain some insight into the structure of the optimal policies, and also, adding multiple servers to the model increases the complexity of the model and make less tractable. In this section, we consider relaxing this assumption by estimating cases with multiple servers. There are two approaches to approximate the results with multiple servers; reducing service times or

number of the casualties proportionally. Mills et al. (2013) reduce service time to estimate multiple servers. For instance, a case with service time of 30 and one server, when having two servers, is assumed to have a service time of 15 with everything else remaining the same. Kamali et al. (2016) show that modifying service times changes the structure of the optimal solution. They propose reducing number of the casualties in each category proportionally to estimate multiple servers. As an example, in a case with one server and 50 and 40 casualties of types 1 and 2, respectively, they propose number of the casualties should be changed to 25 and 20 to approximate the solution of having two servers (see Kamali et al., 2016, for details). Figure 3.9 illustrates the approximation of the results for having one, two, and three servers.



$(n_1, n_2) = (60, 60)$ $(n_1, n_2) = (30, 30)$ $(n_1, n_2) = (20, 20)$

Figure 3.9: Results from Model 1 with equal rewards $((\alpha_1, \alpha_2) = (1.00, 1.00))$, $(s_1, s_2) = (20, 20)$, $(l_1, l_2) = (200, 300)$, and $(d_1, d_2) = (200, 200)$.

Figure 3.9 shows the results for a case with one server, 60 casualties of each type, $(\alpha_1, \alpha_2) = (1.00, 1.00))$, $(s_1, s_2) = (20, 18)$, $(l_1, l_2) = (200, 300)$, and $(d_1, d_2) = (200, 200)$. The results indicates serving a few type 1 casualties, then all type 2 casualties, followed by the remaining type 1 casualties. The next figure $((n_1, n_2) = (30, 30))$ shows the approximated results for two servers. In this case, a larger number of type 1 results are served initially, while the overall structure of the service remains the same. Finally, when having three servers

($(n_1, n_2) = (20, 20)$), most of the type 1 casualties are initially served before switching to type 2 casualties. In general we observe that increasing servers shifts the results towards $S(1, 2)$, which is consistent with the findings of Kamali et al. (2016). This shift occurs due to the decrease in the resources-constrainedness of the problem as more servers become available. This finding indicates that if more resources become available during the relief efforts, it is more likely that assigning them to type 1 casualties is more effective.

3.8 Conclusions

The medical community has previously identified the need for incorporation of several factors such scale of the disaster and availability of resources into the triage process. In this paper, we develop mathematical models to show how scale of the disaster and availability of the resources could affect the outcome of the triage operation. To the best of our knowledge, our model is the first to account for both on-site and off-site casualty deaths. We also provide two simplifications of our general model, one under off-site, and another under on-site death assumption. Results from the analysis of our models provide decision makers with optimal prioritization policies in order to maximize the total survival probabilities. We achieve this with minimal data requirements. Only number of casualties in each category and service times for each casualty type need to be collected in the aftermath of a disaster. Both of these information can be collected rapidly to generate the optimal triage plan.

Our findings verify the impact disaster-specific factors can have on the outcome of the triage process. This is in contrast with the current triage processes in practice such as START that have a static prioritization policy regardless of the scale of the disaster and available resources. We show although more critical-first policies similar to $S(1, 2)$ might be optimal in some cases, in many other, reverse or some other mixed strategies perform significantly better. We develop an easy to implement heuristic for this problem using the insights gained from the structure of the optimal solution, and we show it performs comparable to optimal

methods. We also briefly study the case with multiple servers, and analyze how increasing the resources affects the optimal results.

While our mathematical model incorporates several details of the real world MCIs, there are still other assumptions of the model that can be relaxed. Currently, we are assuming that we have one server for providing care to all the casualties. We can relax this constraint by having multiple servers simultaneously serving casualties. To be closer to practice, we can augment the size of the fleet through time. We can also extend the definition of servers and have hospitals and ambulances. In this research, and to our knowledge in other related works in the literature, proposed policies are for cases with one server, which could be either the transportation vehicle or the receiving hospital. We need to verify how the proposed policies perform in more complex scenarios with multiple hospitals and emergency vehicles. We also assume that all the casualties are available at the beginning of the relief effort, but in many cases, identification of all casualties might take several hours or even days. Stochastic casualty arrival can be incorporated into the model to address this issue and analyze how it affects the optimal policies. Another assumption we made is having one disaster location. In some cases, there are multiple locations hit by a disaster and prioritization and transportation of the casualties should be done simultaneously in multiple locations.

3.9 Appendix

Proof of Proposition 3. Model 2B is a modification of the model introduced in Kamali et al. (2016), with a single server. The definitions of the decision variables and formulation follow:

Decision Variables:

x_{it} 1 if a casualty of type i starts service in time interval t, 0 otherwise, $\forall i = 1, \ldots, P, t = 1, ..., T$

$$\textbf{Model 2B:} \quad \text{Max} \quad \sum_{i=1}^{P}\sum_{t=1}^{T} f_{it}x_{it} \tag{3.10}$$

$$\text{s.t.} \quad \sum_{t=1}^{T} x_{it} = n_i, \quad \forall i = 1, \ldots P \tag{3.11}$$

$$\sum_{i=1}^{P}\sum_{f=0}^{min(s_i-1,t)} x_{i(t-f)} \leq 1, \quad t = 1, \ldots, T \tag{3.12}$$

$$x_{it} \in \{0, 1\}, \ \forall i = 1, \ldots, P, \ t = 1, \ldots, T. \tag{3.13}$$

Objective function (3.10) maximizes total survival probability for all the casualties. Constraint (3.11) guarantees that all the casualties are served. Constraint (3.12) makes the server unavailable during service (i.e., when the ambulance is transporting a casualty). Constraint (3.13) is the binary and non-negativity constraint. We show that the optimal solution to Model 2 is equivalent to that of Model 2B.

Both models maximize the survival probability defined by the parameter f_{it} for all $i \in \{1, 2\}$ and $t \in 1, \cdots, T$. Model 2 serves all casualties by starting from the state with all casualties, $V(Q, 0)$, where $Q = (n_1, \ldots, n_P)$, which is ensured through constraint (3.10) in Model 2B. After each service, Model 2 determines which casualty type to serve and advances time

properly based on the selected casualty type's service time, which is what Model 2B does through the variables x_{it} and constraint (3.12) that ensures one service at a time. These conditions ensure that an optimal solution from Model 2 provides an optimal solution to Model 2B. In addition, Kamali et al. (2016) show that their Model 2B provides optimal solution to Mills et al. (2013) fluid formulation and Sacco et al. (2007) linear program, and hence, Model 2 provides the optimal solution to those as well. □

Proof of Proposition 4. Exponentially distributed survival probabilities are in the form of $f_{it} = e^{-\frac{t}{l_i}}$. For Assumption 3 to hold, first we need to show there is a t_m such that $\alpha_1 f^{\delta}_{1t_m} = \alpha_2 f^{\delta}_{2t_m}$.

$$
\begin{aligned}
\alpha_1 f^{\delta}_{1t_m} = \alpha_2 f^{\delta}_{2t_m} \quad &\rightarrow \quad \alpha_1(e^{-\frac{t+\delta}{l_1}} - e^{-\frac{t}{l_1}}) = \alpha_2(e^{-\frac{t+\delta}{l_2}} - e^{-\frac{t}{l_2}}) \\
&\rightarrow \quad \alpha_1 e^{-\frac{t}{l_1}}(e^{-\frac{\delta}{l_1}} - 1) = \alpha_2 e^{-\frac{t}{l_2}}(e^{-\frac{\delta}{l_2}} - 1) \\
&\rightarrow \quad \frac{e^{-\frac{t}{l_1}}}{e^{-\frac{t}{l_2}}} = \frac{\alpha_2}{\alpha_1}\frac{e^{-\frac{\delta}{l_2}} - 1}{e^{-\frac{\delta}{l_1}} - 1} \\
&\rightarrow \quad e^{-\frac{t}{l_1}+\frac{t}{l_2}} = \frac{\alpha_2}{\alpha_1}\frac{e^{-\frac{\delta}{l_2}} - 1}{e^{-\frac{\delta}{l_1}} - 1} \\
&\rightarrow \quad t(\frac{-1}{l_1} + \frac{1}{l_2}) = \ln(\frac{\alpha_2}{\alpha_1}\frac{e^{-\frac{\delta}{l_2}} - 1}{e^{-\frac{\delta}{l_1}} - 1}) \\
&\rightarrow \quad t_m = \frac{l_1 l_2}{l_1 - l_2}\ln(\frac{\alpha_2}{\alpha_1}\frac{e^{-\frac{\delta}{l_2}} - 1}{e^{-\frac{\delta}{l_1}} - 1}) \qquad (3.14)
\end{aligned}
$$

For t_m showed in Expression 3.14, we have $\alpha_1 f^{\delta}_{1t_m} = \alpha_2 f^{\delta}_{2t_m}$. We also have $\alpha_1 f^{\delta}_{1t} < \alpha_2 f^{\delta}_{2t}$ for $t < t_m$ and the opposite for $t > t_m$. Thus, exponentially distributed survival probabilities with deterministic service times and constant rewards adhere to Assumption 3. □

Proof of Proposition 5. There are two cases in which a death is observed; At the time of casualty's death, l_i, when the observation time is already passed (i.e., $d_i \leq l_i$), or at observation time, d_i, when the casualty has already died (i.e., $l_i < d_i$). Therefore, death

observation distribution is $\max(l_i, d_i)$. Since both survival and observation probabilities follow exponential distribution, we can calculate the expected value of $\max(l_i, d_i)$ as follows:

$$\begin{aligned} E\left[\max(l_i, d_i)\right] + E\left[\min(l_i, d_i)\right] &= l_i + d_i \\ E\left[\max(l_i, d_i)\right] &= l_i + d_i - E\left[\min(l_i, d_i)\right] \\ &= l_i + d_i - \frac{1}{\frac{1}{l_i} + \frac{1}{d_i}} \\ &= l_i + d_i - \frac{l_i d_i}{l_i + d_i} \end{aligned}$$

The $\min(l_i, d_i)$ is a random variable equal to multiplication of survival (with parameter $\frac{1}{l_i}$) and observation (with parameter $\frac{1}{d_i}$) distributions, which follows another exponential distribution with parameter $\frac{1}{l_i} + \frac{1}{d_i}$. □

Proof of Proposition 6. There is at least one time interval we can get to each casualty state in Model 1's state space (due to the probability of observing casualty's eventual death). At each casualty state, there is only one possible path to reach the largest time interval, which is by serving all casualties up to that state (either alive or unobserved dead). If we serve any less casualties, time interval at that state will be smaller, and there is no way of serving more casualties as all have been served. Also, the only path for getting to a largest time interval at a state, is through another largest time interval for a previous state (Since all casualties should be served at each state). As a result, the decision tree of Model 1 containing only the paths leading to the largest time intervals for each casualty state becomes the same as Model 2, in which all casualties are served. Thus, the optimal decisions regarding largest time interval for each casualty state of Model 1, indicate the optimal solution to Model 2. □

Proof of Proposition 7. Since the survival probability never increases, and the number of casualties of each type are known at time zero, the server will not be idle until all casualties

are served. Given n_i casualties requiring a service time of s_i, $i = 1, \ldots, P$, the total time required to serve all casualties is $\sum_{i=1}^{P} n_i s_i$ (assuming $|P|$ is finite). The dynamic program has two state variables, namely Q, and t. Q contains the number of remaining casualties of each type and their initial values are $(n_1, \ldots, n_P)$, and t is bounded by $\sum_{i=1}^{P} n_i s_i$. The running time of Model 1 is bounded by the size of its state space, which is $\prod_{i\ 1}^{P} n_i \sum_{i=1}^{P} n_i s_i$. Hence,

$$V(Q, 0) \in O(\prod_{i\ 1}^{P} n_i \sum_{i=1}^{P} n_i s_i) \sim O(\prod_{i\ 1}^{P} n_i \sum_{i=1}^{P} n_i),$$

which is polynomial, assuming finite number of casualty types, P. □

Proof of Proposition 8. Without loss of generality, we assume we have a partial solution that serves two type 2 casualties and m_1 type 1 casualties. Initially, one casualty type 2 is served, then all m_1 type 1 casualties, and finally the last type 2 casualty. We call this partial solution schedule S. We assume rest of the solution remains the same in all of the following schedules. We denote the sum of survival probabilities generated by schedule S, $TR(S)$. Now, we study the problem in three cases.

CASE 1 ($\alpha_1 f_{1t}^{s_2} > \alpha_2 f_{2t}^{s_1}$): In this case we move the first type 2 casualty to the end before the second one, to form schedule S'. Now, we can calculate the difference in the total survival probabilities of schedules S and S' as follows:

$$
\begin{aligned}
TR(S') - TR(S) &= (\alpha_2 f_{2t+m_1 s_1} - \alpha_2 f_{2t}) + \sum_{i=1}^{m_1} \left(\alpha_1 f_{1t+(i-1)s_1} - \alpha_1 f_{1t+s_2+(i-1)s_1}\right) \\
&= (\alpha_2 f_{2t+m_1 s_1} - \alpha_2 f_{2t}) - \sum_{i=1}^{m_1} \left(\alpha_1 f_{1t+s_2+(i-1)s_1} - \alpha_1 f_{1t+(i-1)s_1}\right) \\
&= \left(\alpha_2 f_{2t+m_1 s_1} + \sum_{i=1}^{m_1-1} (\alpha_2 f_{2t+is_1} - \alpha_2 f_{2t+is_1}) - \alpha_2 f_{2t}\right) \\
&\quad - \sum_{i=1}^{m_1} \left(\alpha_1 f_{1t+s_2+(i-1)s_1} - \alpha_1 f_{1t+(i-1)s_1}\right) \\
&= \sum_{i=1}^{m_1} \alpha_2 \left(f_{2t+is_1} - f_{2t+(i-1)s_1}\right) - \sum_{i=1}^{m_1} \alpha_1 \left(f_{1t+s_2+(i-1)s_1} - f_{1t+(i-1)s_1}\right) \\
TR(S') - TR(S) &= \sum_{i=1}^{m_1} \alpha_2 f^{s_1}_{2t+(i-1)s_1} - \sum_{i=1}^{m_1} \alpha_1 f^{s_2}_{1t+(i-1)s_1} \qquad (3.15)
\end{aligned}
$$

There are m_1 terms in each summation of the equation (3.15). We now compare terms in the summations with each other one by one. Given the assumption for this case ($\alpha_1 f^{s_2}_{1t} > \alpha_2 f^{s_1}_{2t}$), we have $\alpha_2 f^{s_1}_{2t} < \alpha_1 f^{s_2}_{1t}$, $\alpha_2 f^{s_1}_{2t+s_1} < \alpha_1 f^{s_2}_{1t+s_1}$, $\alpha_2 f^{s_1}_{2t+2s_1} < \alpha_1 f^{s_2}_{1t+2s_1}$, and .... Thus, we conclude that $\sum_{i=1}^{m_1} \alpha_2 f^{s_1}_{2t+(i-1)s_1} < \sum_{i=1}^{m_1} \alpha_1 f^{s_2}_{1t+(i-1)s_1}$, which translates into $TR(S') - TR(S) < 0$. As a result, when $\alpha_1 f^{s_2}_{1t} > \alpha_2 f^{s_1}_{2t}$ holds, we are always better off serving type 1 casualties first, and then, type 2 casualties.

CASE 2 ($\alpha_1 f^{s_2}_{1t} < \alpha_2 f^{s_1}_{2t}$): In this case we move the second type 2 casualty to the beginning after the first one, to form schedule S''. Now, we can calculate the difference in the total survival probabilities of schedules S and S'' as follows:

$$
\begin{aligned}
TR(S'') - TR(S) &= (\alpha_2 f_{2t+s_2} - \alpha_2 f_{2t+m_1 s_1+s_2}) + \sum_{i=1}^{m_1} \left(\alpha_1 f_{1t+2s_2+(i-1)s_1} - \alpha_1 f_{1t+s_2+(i-1)s_1}\right) \\
&= \sum_{i=1}^{m_1} \left(\alpha_1 f_{1t+2s_2+(i-1)s_1} - \alpha_1 f_{1t+s_2+(i-1)s_1}\right) - (\alpha_2 f_{2t+m_1 s_1+s_2} - \alpha_2 f_{2t+s_2}) \\
&= \sum_{i=1}^{m_1} \alpha_1 \left(f_{1t+2s_2+(i-1)s_1} - f_{1t+s_2+(i-1)s_1}\right) \\
&\quad - \left(\alpha_2 f_{2t+m_1 s_1+s_2} \sum_{i=1}^{m_1-1} \alpha_2 \left(f_{2is_1+s_2} - f_{2is_1+s_2}\right) - \alpha_2 f_{2t+s_2}\right) \\
&= \sum_{i=1}^{m_1} \alpha_1 \left(f_{1t+2s_2+(i-1)s_1} - f_{1t+s_2+(i-1)s_1}\right) \\
&\quad - \sum_{i=1}^{m_1} \alpha_2 \left(f_{2t+s_1+(i-1)s_1+s_2} - f_{2t+(i-1)s_1+s_2}\right) \\
TR(S'') - TR(S) &= \sum_{i=1}^{m_1} \alpha_1 f^{s_2}_{1t+s_2+(i-1)s_1} - \sum_{i=1}^{m_1} \alpha_2 f^{s_1}_{2t+(i-1)s_1+s_2} \qquad (3.16)
\end{aligned}
$$

Equation (3.16) has a similar structure to equation (3.15) and using a same comparison, we conclude that $TR(S'') - TR(S) < 0$. Therefore, when $\alpha_1 f^{s_2}_{1t} < \alpha_2 f^{s_1}_{2t}$ holds, we are always better off serving type 2 casualties first, and then, type 1 casualties.

CASE 3 ($\alpha_1 f^{s_2}_{1t} > \alpha_2 f^{s_1}_{2t}$ up to point t^* and then $\alpha_1 f^{s_2}_{1t} < \alpha_2 f^{s_1}_{2t}$): This case follows Assumption 3, that at some point in time $\alpha_2 f_{2t}$ starts to grow faster than $\alpha_1 f_{1t}$. In this case, functions are multiplied by constant service times, thus, turning point is not going to be t_m. To study this case, we break it into two time intervals, $[t, t^*]$ and $[t^*, t + m_1 s_1 + 2s_2]$. In the first interval, based on Assumption 3 we have $\alpha_1 f^{s_2}_{1t} > \alpha_2 f^{s_1}_{2t}$. We showed in case 1 that under this setting, we are better off serving type 1 casualties first and then type 2 casualties. In the second interval, we have $\alpha_1 f^{s_2}_{1t} < \alpha_2 f^{s_1}_{2t}$. This is similar to case 2, in which we showed we are better off serving type 2 casualties first, and then type 1 casualties. Combining results from these two intervals, we have a portion of type 1 casualties served first, then all type 2

casualties, and finally rest of the type 1 casualties. Thus, we are better off not interrupting service of type 2 casualties, even if breaks service to type 1 casualties. □

Proof of Proposition 9. To show this property, without loss of generality we show that for any two arbitrary casualties, one type 1 and a type 2, this property holds. These two casualties can be anywhere in the service schedule in any of the servers. We assume there are two partial schedules starting at time t'; S where type 1 casualty is served first and then type 2, and S' with reverse order. In the following, $s_1 = s_2 = s$ and $\alpha_1 = \alpha_2 = 1$.

$$\begin{aligned} TR(S) - TR(S') &= f_{1t'}(f_{2t'+s}) + (1 - f_{1t'})(f_{2t'}) - (f_{2t'}(f_{1t'+s}) + (1 - f_{2t'})(f_{1t'})) \\ &= f_{1t'}f_{2t'+s} + f_{2t'} - f_{2t'}f_{1t'+s} - f_{1t'} \\ &= f_{2t'}(1 - f_{1t'+s}) - f_{1t'}(1 - f_{2t'+s}) \end{aligned} \tag{3.17}$$

In Expression 3.17, $f_{2t'} > f_{1t'}$ and also, $1 - f_{1t'+s} > 1 - f_{2t'+s}$ for any t' and s. Therefore, $f_{2t'}(1 - f_{1t'+s}) > f_{1t'}(1 - f_{2t'+s})$ and $TR(S) - TR(S') > 0$. This shows that when $s_1 = s_2$ and $\alpha_1 = \alpha_2$, using binary comparison we can show we are always better off serving type 1 casualties first, then type 2 casualties. □

Proof of Proposition 10. The model proposed by Jacobson et al. (2012) relies on the memory-less property of the exponential distribution used for lifetimes and service times. □

Proof of Proposition 11. Limiting possible cases for the optimal policy (see Proposition 8) allows us to calculate the total survival probability as a function of switching point, n_1', which is the number of type 1 casualties served before switching to serve type 2 casualties. Expression (3.18) gives the total survival function, $TTR(n_1')$.

$$TTR(n_1') = \alpha_1 \sum_{i=0}^{n_1'-1} f_{1is_1} + \alpha_2 \sum_{i=0}^{n_2} f_{2n_1's_1+is_2} + \alpha_1 \sum_{i=0}^{n_1-n_1'-1} f_{1n_1's_1+n_2s_2+is_1} \tag{3.18}$$

In order to find $n_1'^*$ when survival probabilities follow exponential distribution, we need to replace functions, $f_1(t)$ and $f_2(t)$, in the total survival function (3.18), and find $n_1'^*$ by finding the root to the derivative as in expression (3.9).

$$\begin{aligned} TTR(n_1') &= \alpha_1 \sum_{i=0}^{n_1'-1} (1 - e^{-\frac{is_1}{l_1}}) + \alpha_2 \sum_{i=0}^{n_2-1} (1 - e^{-\frac{n_1's_1+is_2}{l_2}}) + \alpha_1 \sum_{i=0}^{n_1-n_1'-1} (1 - e^{-\frac{n_1's_1+n_2s_2+is_1}{l_1}}) \\ &= \alpha_1 n_1' - \alpha_1 \sum_{i=0}^{n_1'-1} (e^{-\frac{is_1}{l_1}}) + \alpha_2 n_2 - \alpha_2 e^{-\frac{n_1's_1}{l_2}} \sum_{i=0}^{n_2-1} (e^{-\frac{is_2}{l_2}}) \\ &\quad + \alpha_1 n_1 - \alpha_1 n_1' - \alpha_1 e^{-\frac{n_1's_1}{l_1}} e^{-\frac{n_2s_2}{l_1}} \sum_{i=0}^{n_1-n_1'-1} (e^{-\frac{is_1}{l_1}}) \end{aligned}$$

Considering that $\sum_{i=0}^{N} e^{ia} = \frac{1-e^{Na}}{1-e^a}$, we can simplify the $TTR(n_1')$ function further.

$$\begin{aligned} TTR(n_1') &= \alpha_1 n_1 + \alpha_2 n_2 - \alpha_1 \frac{1 - e^{-\frac{n_1's_1}{l_1}}}{1 - e^{-\frac{s_1}{l_1}}} - \alpha_2 e^{-\frac{n_1's_1}{l_2}} \frac{1 - e^{-\frac{n_2s_2}{l_2}}}{1 - e^{-\frac{s_2}{l_2}}} - \alpha_1 e^{-\frac{n_1's_1}{l_1}} e^{-\frac{n_2s_2}{l_1}} \frac{1 - e^{-\frac{(n_1-n_1')s_1}{l_1}}}{1 - e^{-\frac{s_1}{l_1}}} \\ &= \alpha_1 n_1 + \alpha_2 n_2 - \frac{1 - e^{-\frac{n_1's_1}{l_1}}}{1 - e^{-\frac{s_1}{l_1}}} - \alpha_2 e^{-\frac{n_1's_1}{l_2}} \frac{1 - e^{-\frac{n_2s_2}{l_2}}}{1 - e^{-\frac{s_2}{l_2}}} - \alpha_1 e^{-\frac{n_2s_2}{l_1}} \frac{e^{-\frac{n_1's_1}{l_1}} - e^{-\frac{n_1s_1}{l_1}}}{1 - e^{-\frac{s_1}{l_1}}} \end{aligned}$$

In order to find $n_1'^*$, we need to solve $\frac{dTTR(n_1'^*)}{dn_1'} = 0$.

$$\begin{aligned}\frac{dTTR(n_1')}{dn_1'} &= -\alpha_1\frac{s_1}{l_1}\frac{e^{-\frac{n_1's_1}{l_1}}}{1-e^{-\frac{s_1}{l_1}}}+\alpha_2\frac{s_1}{l_2}e^{-\frac{n_1's_1}{l_2}}\frac{1-e^{-\frac{n_2s_2}{l_2}}}{1-e^{-\frac{s_2}{l_2}}}+\alpha_1\frac{s_1}{l_1}e^{-\frac{n_1's_1}{l_1}}\frac{e^{-\frac{n_2s_2}{l_1}}}{1-e^{-\frac{s_1}{l_1}}}\\ &= -\alpha_1\frac{s_1}{l_1}e^{-\frac{n_1's_1}{l_1}}\frac{1-e^{-\frac{n_2s_2}{l_1}}}{1-e^{-\frac{s_1}{l_1}}}+\alpha_2\frac{s_1}{l_2}e^{-\frac{n_1's_1}{l_2}}\frac{1-e^{-\frac{n_2s_2}{l_2}}}{1-e^{-\frac{s_2}{l_2}}}=0\end{aligned}$$

Now we can find $n_1'^*$ from the above.

$$\alpha_1\frac{s_1}{l_1}e^{-\frac{n_1's_1}{l_1}}\frac{1-e^{-\frac{n_2s_2}{l_1}}}{1-e^{-\frac{s_1}{l_1}}}=\alpha_2\frac{s_1}{l_2}e^{-\frac{n_1's_1}{l_2}}\frac{1-e^{-\frac{n_2s_2}{l_2}}}{1-e^{-\frac{s_2}{l_2}}}$$

$$\frac{e^{-\frac{n_1's_1}{l_1}}}{e^{-\frac{n_1's_1}{l_2}}}=\frac{\alpha_2l_1}{\alpha_1l_2}\frac{\frac{1-e^{-\frac{n_2s_2}{l_2}}}{1-e^{-\frac{s_2}{l_2}}}}{\frac{1-e^{-\frac{n_2s_2}{l_1}}}{1-e^{-\frac{s_1}{l_1}}}}$$

$$e^{-n_1's_1(\frac{1}{l_1}-\frac{1}{l_2})}=\frac{\alpha_2l_1}{\alpha_1l_2}\frac{\frac{1-e^{-\frac{n_2s_2}{l_2}}}{1-e^{-\frac{s_2}{l_2}}}}{\frac{1-e^{-\frac{n_2s_2}{l_1}}}{1-e^{-\frac{s_1}{l_1}}}}$$

$$n_1's_1(\frac{1}{l_2}-\frac{1}{l_1})=\ln\left(\frac{\alpha_2l_1}{\alpha_1l_2}\frac{\frac{1-e^{-\frac{n_2s_2}{l_2}}}{1-e^{-\frac{s_2}{l_2}}}}{\frac{1-e^{-\frac{n_2s_2}{l_1}}}{1-e^{-\frac{s_1}{l_1}}}}\right)$$

$$n_1'^*=\frac{l_1l_2}{s_1(l_1-l_2)}\ln\left(\frac{\alpha_2l_1}{\alpha_1l_2}\frac{\frac{1-e^{-\frac{n_2s_2}{l_2}}}{1-e^{-\frac{s_2}{l_2}}}}{\frac{1-e^{-\frac{n_2s_2}{l_1}}}{1-e^{-\frac{s_1}{l_1}}}}\right).\qquad\square$$

Chapter 4

Coordinating regional response following a mass-casualty incident

Abstract

In this paper we study coordinating the response to a mass casualty incident (MCI). An effective and efficient response requires coordination between several entities and to the best of our knowledge, all the related studies in the literature study entities in isolation. We develop a model that incorporates casualty's service order, transportation, and hospital selection. The objective of the model is to maximize the total survival probability for all the casualties. We develop a novel survival probability function that accounts for type of the casualty, level of care during the transportation, travel distance, and receiving hospital's quality of care. We compare the results from our model to the common real world policies and relevant studies in the literature and analyze the performance of our model under various resource settings.

4.1 Introduction

Mass-casualty incidents (MCI) overwhelm the medical resources in a region due to the sudden increase in demand. Quantity and extent of the disasters, both natural man-made, have significantly increased throughout the world in the past decades. The main reasons for such increase are population growth, urbanization, development of advanced technologies, environmental degradation, and global social interdependence (Ginter et al., 2006). An effective response to an MCI requires a coordinated effort between several organizations and an efficient utilization of resources. Involved organizations include local hospitals and care facilities, police departments, fire departments, and other local and national humanitarian organizations (e.g., Red Cross). There could be independent players involved in the rescue effort such as volunteers, but they are often managed by one or more of the involved organizations. Resources refer to the capacity and capability of the involved organization in transferring, accommodating, and providing care for the casualties. For instance, resources in a hospital include nurses, physicians, medical supplies, operating rooms, and beds.

Typically the first step in the response to an MCI following arrival of the emergency personnel is the triage process. Triage refers to categorization of the casualties based on their condition. There are several triage methods developed, including START (Simple Triage and Rapid Treatment), Homebush, Triage Sieve, CareFlite, Sacco Triage Method (STM), Military Triage, SALT (sort, assess, life-saving interventions, treatment and/or transport), and CESIRA, but none of them are agreed upon as a global approach for dealing with MCIs (Lerner et al., 2008). In addition, there are several triage methods specifically developed for pediatric casualties including JumpSTART and Pediatric Triage Tape (PTT). The main difference between these triage methods is the number of casualty categories and how casualties are divided among the categories. Some of these triage methods are widely used for decades now, but there is no or very little evidence on validity and effectiveness of any of these triage methods (Jenkins et al., 2008). Most of these triage methods come with a fixed inherent policy for prioritization of the casualties based on their type from most critical to least, regardless of the characteristics of the disaster and available resources. Several studies in the literature demonstrate that such static policy is not necessarily optimal, and in fact, the worst policy under certain circumstances (Sacco et al., 2005; Jacobson et al., 2012; Mills et al., 2013; Kamali et al., 2016). These studies propose different approaches for generating dynamic ordering policies based on simple attributes of the incident. Majority of these studies only consider service priority, with no transportation and hospital assignment (one exception is Sacco et al. (2005), which considers transportation capacities in a basic form). Moreover, to provide some insight into the optimal strategies, most of these studies analyze simplified cases of the problem with two casualty types and one server, which is usually not the case in reality.

Current triage methods do not provide any frameworks or guidelines following the fixed-prioritization policy for transportation and assignment of casualties to the nearby hospitals. Typically, the incident command manager and other emergency personnel at the disaster location make such decisions in ad-hoc manner. There are three common strategies for assigning casualties to hospitals (Dean and Nair, 2014):

1. Closest-first: start sending casualties to the closest hospital first. Upon reaching capacity, move to the next closest hospital,

2. Farthest-first: start sending casualties to the farthest hospital first. Upon reaching capacity, move to the next farther hospital,

3. Cyclical: send casualties in a round robin manner to the local hospitals.

All of these strategies have a fixed nature and they do not incorporate any information about transportation resources and hospitals such as hospital capacity, beds available, emergency vehicles available, or route duration. There are cases that a hospital is more suited for specific casualty types based on their needs, and thus, hospital assignment needs to consider factors other than distance as well. Another point missing from these static policies is how resources may vary over time. For instance, immediately following a disaster there a limited fleet of emergency vehicles might be available, which could increase gradually when resources are moved to the affected region from the neighboring areas. Same scenario happens in hospitals as they activate their surge capacity following an MCI, but the increase happens progressively, as opposed to immediately. While several studies analyze surge capacity in hospitals and propose conceptual frameworks to address surge capacity in case of an MCI, to the best of our knowledge, none of the studies in this area consider the increase in resources such as surge capacity.

Out of the operations in an emergency response, transportation of casualties has received the least attention. This is partly due to the simpler nature of the transportation decisions; once we know the order of transferring casualties and their receiving hospital, we use one of the available vehicles to transfer the casualty. If no vehicle is available, upon return of one to the disaster site, the casualty is transported. One challenging case is when there are multiple options available for moving casualties, such as ambulances or helicopters. In this case, the decision is typically at the mercy of the incident command managers. Another reason for less focus on transportation of casualties is the extensive body of research in

the area of transportation, of which many cases can be applied to this problem with minor modifications.

As described above, there is a large body of literature studying the disaster response problem, but to the best of our knowledge, all of these studies focus on one specific aspect of the response. Sacco et al. (2005); Jacobson et al. (2012); Mills et al. (2013); Kamali et al. (2016) develop models to study the service priority problem. Sacco et al. (2005) only provides a model along with some results, but the rest provide some insights into the structure of the optimal policies under some simplifying assumptions. Dean and Nair (2014) develop a model to address resource allocation in addition to service order. We discuss this model in details later in this paper. There are also several other studies focusing on surge capacity management (Kanter and Moran, 2007a; Lerner et al., 2008; Hick et al., 2009) and transportation (Castle, 2006), but they either provide a conceptual framework or discuss the issues in the area. While study of each stage is equally important to gain some insight into the structure of the optimal policies, most of the decision in response to an MCI affect each other and in other words, we cannot find the best policy by studying pieces in isolation. For instance, service order could affect the decision regarding the receiving hospital, which subsequently could determine the vehicle type to transfer the casualty. Thus, in order to fully understand the structure of the optimal strategies for the response to an MCI, we need to incorporate all the mentioned elements and study how they interact with each other.

Our main contribution in this paper is incorporation of service order, transportation, hospital assignment, and capacity management in response to an MCI and analyze optimal strategies. Our novel approach addresses resources variability over time and reusability of resources. This paper is structured as follows: In §2 we define the problem under study state the underlying assumptions we are using, and describe our proposed formulation for this problem in §3. We perform a numerical analysis in §4 and compare our results to that of related papers in the literature. We provide implementation guidelines in §5, and finally in §6 we provide concluding remarks and future areas of improvement.

4.2 Problem Definition

Consider a mass-casualty incident with a certain number of casualties categorized into several types, based on the criticality of their condition and survival chance. We assume we have the number of casualties and their types in advance prior to beginning service (i.e., no future arrivals). Also, we assume casualties' type does not change over the course of service. We use a time-dependent non-increasing survival probability to capture the degradation in casualties' condition while waiting for service. Throughout this paper we assume casualty types are sorted based on criticality, with the first casualty type being the most critical (i.e., smallest survival probability) and the last being the least critical (i.e., largest survival probability). Without loss of generality, we divide the time horizon into equal time intervals. We assume there is a fleet of emergency vehicles available that is divided by type. Each type has a specific speed factor, which determines the travel time of that type. As mentioned earlier, immediately following a disaster there could be a limited number of emergency vehicles available to serve the casualties, but this number can gradually increase as other vehicles arrive from surrounding areas. To capture this variability, it is assumed that number of the total vehicles of each type is given at each time interval (this can be estimated by the hospitals in a region). The travel time to each hospital, loading, and unloading times are known in advance and following moving a casualty to a receiving hospital, vehicles return to the disaster location to serve the remaining casualties.

Upon arrival of casualties at the hospitals, life-saving interventions are performed to stabilize the casualties. Then, casualties are moved to other sections or wards where they are monitored and stay longer for full recovery. Based on these two stages of the care, we define two types of capacities for each hospital. First, the capacity for immediate service and performing life-saving interventions, which requires physicians, nurses, operating rooms, medical supplies, and equipment. After stabilization, the main requirement is a bed for casualties to stay, while the staff monitor their recovery. Both of these capacities can increase over time as hospitals call in more staff, access their backup supplies, and activate their surge

beds. There multiple surge capacity requirements, thus hospitals have an estimation of how resources increase over time and we assume we have the total amount of these resources at each point in time. Based on the criticality of their condition, we assume each casualty type occupies a certain amount of resources for a certain number of time intervals. After that time, resources become available again to serve other casualty waiting for service (if any). In the next section, we introduce our proposed formulation for the described problem.

4.3 Model Formulation

To model this problem, we assume we have a set of casualty types, C, a set of receiving facilities (e.g., hospitals), H, and a set of emergency vehicle types, V. The time horizon for service to all the casualties is divided to T time intervals, starting with time 0 and ending at T. We assume all the casualties are available at time 0 and there is no future arrival of casualties. The service is provided in a non-preemptive manner and continues until all casualties are served. A non-increasing survival probability function is defined for each casualty type i denoted by f_{chvt}, which is calculated based on the expected lifetime of that casualty type for all $c \in C$, receiving hospital $h \in H$, emergency vehicle $v \in V$, and $t \in \{0, \ldots, T\}$. Throughout the rest of this paper, we assume casualty types are sorted from type 1 being the most critical to type $|P|$ being the least critical ($f_{1hvt} \leq f_{2hvt} \leq \ldots \leq f_{|P|hvt}, \forall h \in H, v \in V, t \in \{1, \ldots, T\}$). The complete list of parameters and decision variables in formulating this problem are as follows.

Sets and Parameters:

C	set of casualties sorted based on criticality
H	set of participating care facilities
V	set of vehicle types
T	number of time intervals in the time horizon
n_c	number of casualties of type c at the beginning of time interval 1, $\forall c \in C$

f_{chvt} survival probability for casualty type c transported at time t using a vehicle of type v to the receiving hospital h, $\forall c \in C, h \in H, v \in V, t = 1, \ldots, T$

d_{hv} number of time intervals required to travel from disaster location to care facility h using a vehicle of type v, $\forall h \in H, v \in V$

λ_c time intervals required to load/unload casualties on/from transportation vehicles, $\forall c \in C$

$\lambda_{\min}$ minimum loading/unloading value, $\lambda_{\min} = min_{c \in C}\{\lambda_c\}$

m maximum vehicles allowed in the disaster location for loading casualties

α_{vt} total number of vehicles of type v in the fleet at time t, $\forall v \in V, t = 1, \ldots, T$

r_c resource requirements for a casualty of type c, $\forall c \in C$

τ_{ch} time intervals a casualty of type c needs resources at hospital h upon arrival, $\forall c \in C, h \in H$

ρ_{ht} total amount of shared care resources in hospital h at time t, $\forall h \in H, t = 1, \ldots, T$

β_{cht} total long-term beds for casualties of type c in hospital h at time t, $\forall c \in C, h \in H, t = 1, \ldots, T$

Decision Variables:

x_{chvt} number of the casualties of type c transferred to hospital h by vehicles of type v at time t, $\forall c \in C, h \in H, v \in V, t \in \{1, \ldots, T\}$

y_{chvt} number of the casualties of type c arrived at hospital h by a vehicle of type v at time t, $\forall c \in C, h \in H, v \in V, t \in \{1, \ldots, T\}$

Given the defined parameters and decision variables, we present our formulation for the problem, Model 1.

Model 1: max $$\sum_{c\in C}\sum_{h\in H}\sum_{v\in V}\sum_{t=1}^{T} f_{chvt}x_{chvt} \tag{4.1}$$

s.t. $$\sum_{h\in H}\sum_{v\in V}\sum_{t=1}^{T} x_{chvt} = n_c, \quad \forall c \in C \tag{4.2}$$

$$\sum_{c\in C}\sum_{h\in H}\sum_{u=1}^{\min(t,2\lambda_c+2d_{hv}+1)} x_{chv(t-u+1)} \leq \alpha_{vt}, \quad \forall v \in V, t = 1,\ldots,T \tag{4.3}$$

$$\sum_{c\in C}\sum_{h\in H}\sum_{v\in V}\sum_{u=\max(1,t-\lambda_c)}^{t} x_{chvu} \leq m, \quad \forall t = 1,\ldots,T \tag{4.4}$$

$$y_{cht} = \sum_{v\in V} x_{chv(t-2\lambda_c-d_{hv})}, \quad \forall c \in C, h \in H, v \in V, t = 2\lambda_c + d_{hv} + 1,\ldots,T \tag{4.5}$$

$$\sum_{c\in C} r_c \sum_{u=1}^{\min(t,\tau_{ch}+1)} y_{ch(t-u+1)} \leq \rho_{ht}, \quad \forall h \in H, t = 2\lambda_{\min} + d_{hv},\ldots,T \tag{4.6}$$

$$\sum_{u=1}^{t} y_{chu} \leq \beta_{ch(t+\tau_{ch})}, \quad \forall c \in C, h \in H, t = 2\lambda_c + d_{hv},\ldots,T-\tau_{ch} \tag{4.7}$$

$$x_{chvt}, y_{cht} \geq 0, \quad \forall c \in C, h \in H, v \in V, t = 1,\ldots,T \tag{4.8}$$

Objective function (4.1) maximizes the total survival probabilities across all the casualties. Constraint (4.2) makes sure that all the casualties are served during the time horizon. Constraint (4.3) guarantee that only available vehicles are used at each time interval. This constraint does not allow previously assigned vehicles to be used again until their service is finished and they are back at the disaster location. Constraint (4.4) assures that no more than the maximum m allowed vehicles are loaded at any time interval in the disaster location. Constraint (4.5) defines the relationship between x and y decision variables. Constraint (4.6) tracks the resource requirements of the assigned casualties at the receiving facility and keeps the resources occupied for the duration of their service. Constraint (4.7) asserts number of the casualties released from immediate care are bounded by the number of long-term beds available for that type in each hospital. Finally, constraint (4.8) makes sure all variables are non-negative.

Among the studies in the area of MCI response, two of the works in the literature have more relevance to the problem we study here. Sacco et al. (2005) develop a simple model called Sacco Triage Method (STM) with the objective of maximizing total survival probability, while transportation resources are limited. They define a resource level for each time period and limit the transportation resources to that. Dean and Nair (2014) extend the model developed by Sacco et al. (2005) to consider resource allocation in addition to service order in their proposed approach, Severity-Adjusted Victim Evacuation (SAVE). They address reusability of both transportation resources and beds in receiving hospitals in their model, but they start with a fixed number of each and keep reusing those. They do not allow gradual increase over time, which happens in practice as more resources are added to the relief operations over the short period of time following the MCI. The capability of our model to consider both reusability and augmentation of resources allows us to generate more effective and realistic policies, as we show in our numerical analysis. Another major difference between our model and SAVE is the casualties' survival probabilities. As explained later, we use a three-piece survival probability, which accounts for the risk exposed to the casualties during the transportation process, in addition to the risk on the disaster location before receiving service. We also add a hospital quality factor that affects the overall survival probability. Moreover, we break hospital resources into shared emergency room resources and casualty type dependent long term beds. This allows us to more accurately model the real world. Dean and Nair (2014) only define casualty type dependent hospital beds, and they reuse them for all the casualties, while skipping emergency room processing in SAVE. Typically when casualties are stabilized following their arrival at the receiving hospital, they are moved to intensive care units or other types of beds based on their condition and stay there for a longer period of time, which is beyond the scope of the MCI relief efforts. We also incorporate several other characteristics of the problem into our model, such as loading and unloading times, and a limit on number of the vehicles that can arrive at a hospital at the same time. As we show later in the numerical analysis, not considering the limit on the simultaneous casualties being processed can cause some issues in SAVE results.

4.4 Survival Data

Majority of the parameters and data needed to solve this problem are relatively straight forward to obtain. Several pieces of information can even be acquired in advance, such as number of the hospitals in the region, their capacity, and size of the emergency vehicles fleet, while the rest is gathered upon assessment of the MCI by the emergency personnel (e.g., number of the casualties and travel time to regional care facilities). Some of the parameters including the rate in which emergency vehicles and care resources increase can be estimated by the experts and hospital managers, and later we analyze the sensitivity of the results to these parameters. Among the parameters, survival probability is more challenging to obtain. There are few studies in the literature that discuss the survival probabilities of MCI casualties. Sacco et al. (2005, 2007) generate survival probability for 13 casualty types based on the initial condition from the large set of trauma data. Mills et al. (2013) develop a scaled log-logistics distribution-based function from the Sacco et al. (2005) survival curves, and propose five set of parameters for a 2 type classification of the casualties. (Kamali et al., 2016) discuss some issues with this proposed function including the scaling of the survival values. Another distribution used for survival probabilities is the Exponential distribution (Jacobson et al., 2012; Kamali and Bish, 2016). Our developed model allows any distribution for survival probabilities, but we use Exponential distribution-based as explained below throughout the rest of this paper.

The survival probability function, f_{chvt}, takes into account multiple factors including: time spent waiting for service, emergency vehicle selection, and hospital selection. Initially, a casualty's survival follows the Exponential distribution, while they are waiting for service. We refer to this portion of the function as the base survival probability, which is solely dependent on type of the casualty. Assuming a casualty has an expected lifetime of l_c, the base survival probability is calculated as $e^{\frac{-t}{l_c}}$, for all $c \in C,\ t = 1, \ldots, T$. Once the casualty is assigned to an emergency vehicle for transportation, the base survival probability is realized. Beginning transportation until arrival at the receiving hospital, the casualty's

survival probability still decrease, but potentially depending on the level of care provided in the emergency vehicle, it decreases at a lower rate. To capture this decrease, we use the Exponential distribution with a larger parameter, $l_c + \gamma_v$, in which γ_v is the expected number of time periods added to the casualty's lifetime due to the better quality of care during transportation in a vehicle of type $v \in V$. Assuming the transportation is started at time t, decrease in the survival probability during the travel time (d_{hv}) is calculated as $e^{\frac{-d_{vh}}{l_c+\gamma_v}}$ for all $c \in C, h \in H, v \in V, t = 1, \ldots, T$. Improvement in the expected lifetime due to the care provided in the emergency vehicle should be non-negative, $\gamma_v \geq 0$, where value 0 indicates no change in survival probability (e.g., emergency vehicle provides no care), and the larger values indicate better level of care. In order to capture the capabilities of the receiving hospital for each casualty type, we consider a hospital care quality factor, ω_{ch}, which has the maximum value of 1 for the best level of care and lower values for lower levels of care ($0 \leq \omega_{ch} \leq 1$). Based on the given description, we use the following functional form for survival probability in this paper:

$$f_{chvt} = \omega_{ch} e^{\frac{-t}{l_c}} e^{\frac{-d_{hv}}{l_c+\gamma_v}}, \ \forall c \in C, h \in H, v \in V, t = 1, \ldots, T \tag{4.9}$$

Survival probability values are calculated a priori for each combination of casualty type, transportation time, vehicle type, and receiving hospital, and are used in the model as parameters. Figure 4.1 illustrates two sample survival probability curves considering transportation compared to one without any transportation (i.e., transportation with no care provided). The casualty's expected lifetime is 200, $\gamma_{\text{ALS}} = 100$, and $\gamma_{\text{BLS}} = 050$. In the figure, the solid (black) curve shows the base survival probability with mean 200 time units without any transportation provided (or with transportation on a vehicle with $\gamma_v = 0$). In both cases transportation starts at time 100, in one case with an ALS ambulance, highlighted with (green) dashed line, and the other with a BLS ambulance highlighted with (red) double-line. In both scenarios survival remains greater than the base value. To further demonstrate the effect of these factors on overall survival probability of casualties, consider a case with

two hospitals, a farther level I trauma center ($\omega_1 = 1$), 40 time intervals away, and a closer level III trauma center ($\omega_2 = 0.9$), 10 time intervals away. When we have a type 1 casualty in a critical condition with expected lifetime of 200 time intervals, upon availability of an emergency vehicle, emergency personnel are faced with a decision of sending the casualty to the farther hospital with a higher quality of care, or the closer one with lower quality of care. We assume an ALS ambulance with care level $\gamma_{\text{ALS}} = 100$ (as shown in Figure 4.1) becomes available at time 100. In this case $f_{1,1,\text{ALS},100} = 1 \times (0.607 \times 0.875) = 0.531$ and $f_{1,2,\text{ALS},100} = 0.9 \times (0.607 \times 0.967) = 0.528$ (values can be read from Figure 4.1). Thus, in this case, the casualty has a higher chance of survival is transferred to the farther level I trauma center. Now, assume we have the same case, but instead of an ALS, a BLS becomes available. In this case, $f_{1,1,\text{BLS},100} = 1 \times (0.607 \times 0.852) = 0.517$ and $f_{1,2,\text{BLS},100} = 0.9 \times (0.607 \times 0.959) = 0.524$. Unlike the case with ALS, in this case we are better off sending the casualty to the closer hospital, which has a lower quality of care. This simple example highlights the importance of a model that incorporates all these factors into consideration, in order to generate effective and high quality results.

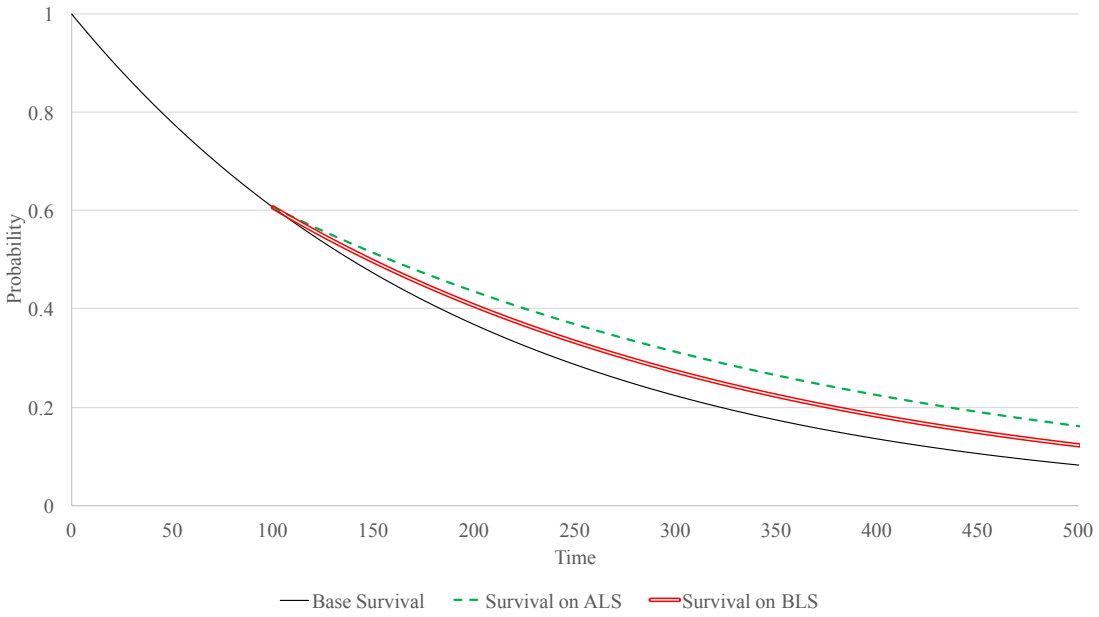


Figure 4.1: Survival probability without and with transportation at times 100 and 200

4.5 Numerical Analysis

In this section we analyze various cases using our proposed formulation. Throughout our analysis we assume there are 25 casualties of each type ($n_c = 25$), and survival probabilities are calculated using Expression 4.9 with expected lifetimes (l_c) of 100, 150, 300, and 350 for each casualty type in order. We assume casualty types 1 and 3 are adults categorized as immediate and delayed based on START method, and types 2 and 4 are pediatrics categorized as immediate and delayed based on JumpSTART method. Since pediatric casualties have a higher resilience and ability to recover, they have a slightly higher expected lifetime than their adult counterparts. One ambulance type is available, with added expected lifetime of 50 ($\gamma_1 = 50$). We assume it takes five time intervals to travel to the first hospital ($d_{1,1} = 5$), and ten to the second ($d_{2,1} = 10$), and two time interval for loading or unloading casualties ($\lambda_c = 2$). We assume the closer hospital ($h = 1$) is a either a level II or III trauma center with ability to provide prompt assessment, surgery, intensive care, and stabilization of inured casualties, and the farther hospital ($h = 2$) is a level I trauma center with the ability to provide total care for every aspect of injury, including to pediatric casualties. Therefore, we assume the care quality factor is $\omega_{1,1} = 0.9$ and $\omega_{3,1} = 1$ for adult casualties in the closer hospital and $\omega_{2,1} = 0.85$ and $\omega_{4,1} = 0.9$ for pediatrics. The farther hospital has the highest care quality factor for all casualty types ($\omega_{c,2} = 1$ for all $c \in C$). The available number of ambulances is set to three, which remains constant during the service period ($\alpha_{1t} = 3$), and the maximum number of allowed vehicles at the disaster location at the same time is set to two ($m = 2$). This makes transportation resources the main constraint in providing service. Resource requirements for casualty types are 4, 3, 2, and 1 in order ($r_c = 5 - c$, for $c \in \{1, \ldots, 4\}$), and the number of time intervals resources remain occupied by each casualty type in both hospitals is 4, 3, 2, and 1, respectively ($\tau_{ch} = 5 - c$). Total amount of resources in both hospitals is constant through time and equal to 20. The number of long-term beds available in both hospitals is also constant and equal to 50 ($\beta_{cht} = 25$). Table 4.1 shows a summary of hospital-related parameters.

Table 4.1: Hospital-related parameters used in the numerical study

Parameter	Hospital	Casualty Type				
		All	1	2	3	4
Initial total shared	1	20				
care resources ($\rho_{h,0}$)	2	20				
Resource requirement (r_c)	1,2		4	3	2	1
Resource time (τ_{ch})	1,2		10	8	6	4
Initial beds ($\beta_{c,h,0}$)	1,2		25	25	25	25
Care level (ω_{ch})	1		0.9	0.85	1	0.9
	2		1	1	1	1

In our numerical analysis, we set number of the beds for each casualty type in each hospital equal to the total number of casualties of that type available, as our purpose is to make sure bed limitations are not a determining factors in hospital selection. For instance, if there are 25 type 1 casualties available, and there are 5 type 1 beds in hospital 1 and 30 in hospital 2, majority of type 1 casualties are forced to be transferred to the hospital 2. While this scenario could occur in real world and our model has the ability to address it, in this section we want to analyze the results when there are no forced limitations. In the rest of this section we study how each element of the problem affects the results in isolation.

4.5.1 Casualty Types

First, we compare the optimal results from our model to the real-world and other simple static policies. We define two basic strategies regarding service order to casualties: $S(1, |P|)$, which refers to starting service to most critical casualty type first, and continue in order and finally serve the least critical casualty type, and $S(|P|, 1)$, which is the reverse order by starting service to the least critical casualty type and continue in order of increasing criticality. We compare the optimal solution generated by our model with $S(1, |P|)$ and $S(|P|, 1)$ under two scenarios with two and four casualty types, one vehicle type, and two hospitals. In the scenario with two casualty types, we assume we only have adults (types 1 and 3 as described earlier), and in the second scenario, all four casualty types are available.

Under the fixed policies, we use three strategies for hospital assignment as mentioned in the literature: assignment to closest hospital with available resources, farthest, and handling hospital assignment using our proposed model. The purpose of the last strategy is to measure the effect of hospital assignment using our proposed model on the total survivors of an MCI. Before discussing the comparison of the results for the described setting, we present the optimal solution from our model for the two mentioned scenarios in Table 4.2.

Table 4.2: Optimal solution from Model 1 for fixed number of vehicles, hospital resources, and long-term beds with two and four casualty types

Scenario	Optimal Solution
1	(1,1,20) - (3,1,25) - (1,1,1) - (1,2,4)
2	(1,1,8) - (2,1,8) - (3,1,25) - (4,1,22) - (4,2,3) - (2,1,12) - (2,2,5) - (1,1,14) - (1,2,3)

Results in Table 4.2 show the order in which casualties are served along with the receiving hospital. Each expression has three terms: casualty type, receiving hospital, and number of the casualties of the mentioned type transferred to the mentioned receiving hospital. In all cases results show a mixed strategy for service, and under no scenario we have a simple $S(1, |P|)$ or $S(|P|, 1)$ strategy. Although service starts with type 1 casualties in both scenarios, as we increase number of the casualties, less type 1 casualties are served initially. Most of the casualties are transferred to the closer hospital, except for some casualties of types 1, 2, and 4 in later times, that are transferred to hospital 2 (i.e., farther). One reason for the later transfer to the level I trauma center is the slower decrease in the survival probability, which justifies the higher level of care gained with a longer travel time. In general, results indicate that as criticality (casualty to resources ratio) increase, priority shifts to less critical casualty types. Table 4.3 show the comparison of the results from $S(1, |P|)$, $S(|P|, 1)$, and Model 1 under the two defined scenarios.

Table 4.3 shows the optimal results from Model 1, along with the results from $S(1, |P|)$ and $S(|P|, 1)$ under three receiving hospital assignment policies: optimal (determined by the model), closest hospital first (CHF), and farthest hospital first (FHF). Columns indicate expected number of survivors from each casualty type and total expected survivors. Results

Table 4.3: Comparison of the results from Model 1, $S(1, |P|)$, and $S(|P|, 1)$ with fixed vehicles, hospital resources, and long-term beds

Method	Hospital Assignment	Expected Survivors				
		Type 1	Type 2	Type 3	Type 4	Total
		Scenario 1				
$S(1, \|P\|)$	Optimal	13.10	-	13.54	-	26.64
$S(1, \|P\|)$	CHF	13.10	-	13.54	-	26.64
$S(1, \|P\|)$	FHF	10.88	-	9.06	-	19.94
$S(\|P\|, 1)$	Optimal	3.76	-	20.52	-	24.27
$S(\|P\|, 1)$	CHF	3.74	-	20.52	-	24.26
$S(\|P\|, 1)$	FHF	1.36	-	18.12	-	19.49
Model 1	Optimal	12.05	-	14.72	-	26.77
		Scenario 2				
$S(1, \|P\|)$	Optimal	13.10	6.18	9.00	6.67	34.95
$S(1, \|P\|)$	CHF	13.10	6.18	9.00	6.52	34.80
$S(1, \|P\|)$	FHF	10.75	3.54	4.73	3.30	22.32
$S(\|P\|, 1)$	Optimal	0.27	2.69	13.29	19.04	35.29
$S(\|P\|, 1)$	CHF	0.29	2.67	13.30	18.97	35.24
$S(\|P\|, 1)$	FHF	0.02	0.86	9.05	18.92	28.85
Model 1	Optimal	6.17	5.81	15.73	10.68	38.38

from Model 1 lead to highest total expected survivors in both scenarios, with a higher difference in the scenario with greater number of casualties. Model 1 results are closer to $S(1, |P|)$, and as we increase the number of casualties, the gap between Model 1 and static policies increase. Among $S(1, |P|)$ and $S(|P|, 1)$, optimal hospital assignment results in selection of the closest hospital for all casualties under first scenario, and under the second scenario, except few casualties, the rest are transferred to the closest hospital. This is aligned with the earlier observation, which is the justification for transferring casualties to the farther hospitals as rate of decrease in survival probability gets slower. As mentioned earlier, as we increase number of the casualties (and casualty types), most critical casualties are affected more than others. Service is delayed to them as relatively small number of resources can save more lives if assigned to other casualty types with a greater survival probability. Next, we analyze how different number of vehicles affect the results from Model 1.

4.5.2 Vehicles

Table 4.4: Comparison of the results from Model 1 with different number of vehicles

	Constant					Gradual Increase				
Vehicles	Type 1	Type 2	Type 3	Type 4	Total	Type 1	Type 2	Type 3	Type 4	Total
1	0.00	0.04	14.37	4.78	19.20	0.00	0.04	14.37	4.78	19.20
2	3.23	2.03	16.11	9.08	30.45	1.62	2.03	16.11	9.08	28.84
3	6.17	5.81	15.73	10.68	38.38	3.02	4.27	15.98	10.79	34.07
4	9.09	9.49	15.16	11.30	45.04	3.22	7.59	15.16	11.27	37.24
5	13.81	10.35	14.82	11.73	50.71	4.46	8.28	14.82	11.55	39.10

Table 4.4 presents the results from Model 1 with different number of vehicles. The setting under which the model is solved is similar to that of Table 4.3, with four casualty types. For each number of vehicles we have defined two scenarios, one where all the vehicles are available initially and the quantity remains constant throughout the relief effort, and another when the service initially begins with one vehicle, and one more added every 30 time intervals until reaching the maximum for that case. Comparison of the results from Table 4.4 shows that gradual increase in number of the vehicles noticeably affects the total expected survivors, especially in cases with more vehicles. This is due to the initial difference in number of the vehicles, which becomes greater when maximum number of the vehicles is larger. We observe that when we increase size of the fleet gradually, the most critical casualty type is affected the most, which is due to the increase in the casualty to resources ratio and their service is delayed after all the less critical casualty types. To further explain the difference in the structure of the optimal solution under the constant and gradual increase policies, we have listed the optimal solution for the case with five vehicles under both policies here:

- Constant: (1,1,20) - (2,1,20) - (3,1,25) - (2,1,1) - (4,1,6) - (2,1,3) - (4,2,19) - (2,2,1) - (1,2,5)
- Gradual: (1,1,6) - (2,1,14) - (3,1,25) - (4,1,12) - (2,1,9) - (4,2,13) - (2,2,2) - (1,1,14) - (1,2,5)

In this case, not only the gradual increase in the fleet changes the service order, but it also alters the receiving hospital for the casualties. For 5 ambulances, under constant scenario 25 casualties are transferred to hospital 2, instead of 20 under gradual increase, as under fewer resources there is less justification to drive longer for slightly better service. Increase in the number of vehicles has a diminishing effect on the total expected survivors, especially when it increases gradually. In that case, time that vehicles reach their maximum values tends to go beyond the relief effort horizon and increase after a certain time interval does not have any effect on the results. These results show the importance of considering variability in number of the vehicles and their gradual increase following an MCI, as it can have a noticeable effect on the outcome of the relief efforts. To the best of our knowledge, Model 1 is the only formulation that considers this variability. Next, we study the effect of multiple vehicle types on the results.

Table 4.5: Comparison of the results from Model 1 with different types of vehicles

	Expected Survival				
Vehicles	Type 1	Type 2	Type 3	Type 4	Total
4 BLS	9.02	9.44	15.14	11.29	44.89
4 ALS	9.09	9.49	15.16	11.30	45.04
2 BLS, 2 ALS	9.06	9.46	15.15	11.29	44.97
2 BLS, 1 ALS, 1 Helicopter	13.76	10.37	14.81	11.69	50.63

Table 4.5 compares the results with different types of vehicles under four scenarios: 4 basic life support (BLS) ambulances, 4 advanced life support (ALS), 2 BLS and 2 ALS, and 2 BLS, 1 ALS, and a helicopter. We assume the increase in expected lifetime for each vehicle type is $\gamma_{\text{BLS}} = 20$, $\gamma_{\text{ALS}} = 50$, and $\gamma_{\text{Helicopter}} = 50$. These vehicles have different speeds, and we assume BLS and ALS have the same speed, for which $d_{1,v} = 5$ and $s_{2,v} = 10$ for $v \in \{\text{BLS}, \text{ALS}\}$, and $d_{1,\text{Helicopter}} = 2$ and $d_{2,\text{Helicopter}} = 4$. There is a slight improvement from 4 BLS to both 2 BLS and 2 ALS, and 4 ALS. Also, among these three scenarios, structure of the service almost remains the same. On the other hand, when one ALS is replaced with a helicopter with faster service times, we observe noticeable improvement in the results. The change affects more critical casualty types, especially type 1, which has the largest increase

in expected survivors. This result highlights the importance of speed of offering service, especially for more critical casualty types. Next, we analyze how shared emergency room resources can affect the total expected survivors.

4.5.3 Hospital Resources

Table 4.6 compares the results with different amount of resources in hospitals under two policies: constant resources and when resources increase gradually. The setting under which the results are generated is similar to that of Table 4.3 with four casualty types. For cases in which resources become available gradually, resources start with one unit and reach their maximum value in 120 time intervals in a linear fashion. We observe that increasing resources improves the total expected survivors with a diminishing effect (similar to increase in number of the vehicles). When all resources are available initially and remain constant, under the scenario with 20 resource units at each hospital, results do not differ when we increase the resources to 30 units. Similar to other presented cases, gradual increase in amount of resources increases the criticality as less resources are available initially, and it hurts type 1 casualties more than other types, as their service tends to be delayed. For instance, when we decrease resources from 20 units available constantly to 5 units, total expected survivors decrease from 38.38 to 37.28 (around 3% decrease), but expected type 1 survivors decrease from 6.17 to 1.12 (around 82% decrease). This shows while the overall outcome might change slightly as hospital resources vary, structure of the solution could change significantly. Results from most of these scenarios with gradual increase of the resources have a mixed structure

Table 4.6: Comparison of the results from Model 1 with variable hospital resources

	Constant					Gradual Increase				
Resources	Type 1	Type 2	Type 3	Type 4	Total	Type 1	Type 2	Type 3	Type 4	Total
5	1.12	7.37	17.53	11.26	37.28	0.18	3.65	17.72	14.35	35.91
10	2.42	9.70	15.47	10.57	38.16	0.24	6.98	16.64	12.77	36.63
20	6.17	5.81	15.73	10.68	38.38	0.23	10.48	15.50	11.13	37.33
30	6.17	5.81	15.73	10.68	38.38	3.14	7.78	15.55	11.10	37.57

unlike previous cases. There are many switches from one casualty type to another.

These results confirm the importance of considering variability in resources. Most of the models in the literature consider a fixed number of vehicles or hospital resources that are available initially and remain constant. There are two exceptions in the literature, Sacco et al. (2005) and Dean and Nair (2014). Next, we analyze the results form our model in comparison with other models in the literature.

4.5.4 Comparison with SAVE

To compare the results from our model to that of Dean and Nair (2014)'s SAVE, we solve Dean and Nair (2014)'s model with the shared set of parameters as defined above. These parameters include number and type of casualties, travel times and number of the emergency vehicles, number of the hospitals and bed types within hospitals, and amount of time hospital resources remain occupied by each casualty type. Then, we use the generated solution and calculate the total survival probabilities based on the survival probability function 4.9. Table 4.7 shows the comparison of the results. In terms of total expected survivors, both models generate relatively similar results, with Model 1 performing slightly better. On the other hand, structure of the results are different, especially for types 1 and 2 with smaller number of the vehicles. Model 1 serves larger number of type 1 casualties than SAVE model, while SAVE serves larger number of type 2 casualties. The same pattern exists between types 3 and 4 casualties, with Model 1 serving more type 3 and less type 4, but the magnitude is smaller than types 1 and 2.

Table 4.7: Comparison of the results from Model 1 and SAVE with different number of vehicles

	Model 1					SAVE				
Vehicles	Type 1	Type 2	Type 3	Type 4	Total	Type 1	Type 2	Type 3	Type 4	Total
2	6.14	5.38	15.91	10.67	38.10	3.51	10.57	14.15	9.66	37.89
3	12.45	8.62	15.11	11.54	47.72	10.54	11.98	14.13	10.87	47.52
4	16.82	11.48	14.87	12.45	55.61	16.82	11.48	14.87	12.05	55.21

While the results in Table 4.7 are relatively close, this does not reflect the full capability of Model 1, as several of its advantages cannot be compared to SAVE. For instance, as mentioned earlier, SAVE only considers beds as the hospital resources. The more critical hospital resource in case of an MCI is emergency room resources, which is shared by all the incoming casualties. Bed assignment often is for a longer period of time, which is beyond the scope of a disaster response. Another example is the limit on number of the vehicles that can arrive at a hospital, which SAVE does not account for. If we have a case with 50 casualties and 50 emergency vehicles, SAVE model transfers all the casualties to a receiving hospital at once, which is impractical. Transferring such large number of casualties at once requires a large number of personnel on disaster location to load the casualties on the vehicles, and also large number of staff to unload and care for the casualties at the receiving hospital. This scenario can potentially create another issue in SAVE, since their objective function only considers the survival probability at the time a casualty is assigned to a vehicle. In this case, there is no difference for the model for transferring casualties to a closer versus a farther one. There are not more casualties left waiting for service, thus, the travel distance is not involved anymore. We overcome this issue by including the travel distance and essentially hospital selection as prat of our survival probability function.

4.6 Conclusions

Different stages of the response to an MCI have studied in isolation in the literature, but to the best of our knowledge, this work is the first to study all phases together. We develop a model that incorporates several factors affecting the outcome of the response to an MCI, including number of the emergency vehicles and their types, number of the hospitals in the region, quality of their care, amount of emergency resources they have, and number of long-term beds they have for each casualty type. Our novel approach to generate survival probabilities associated with each casualty type, combine multiple aspects of service to the casualty, including wait time, type of the emergency vehicle used, distance to the receiving

hospital, and quality of care in the receiving facility. Examining all these elements allow us generate more accurate survival probability, which reflect the unique method of service to each casualty. As shown in the numerical analysis section, considering all these factors allow us to generate results that outperform current fixed strategies and also other proposed approaches in the literature.

Our developed model has many parameters, but most of which can be obtained in advance of solving. Organizations involved in disaster response can estimate types and number of the emergency vehicles, hospital resources and capacities, and emergency personnel serving casualties on site in collaboration prior to a disaster. In addition, many of the involved organizations often participate in emergency drills and simulate various types of MCIs. During these events they can gather better estimation of resources and other-time related parameters such as loading and unloading times. During an MCI, incident command manager only needs to obtain number of casualties of each type and distance to nearby care facilities, and using the service schedule generated by the model, service can be provided to all the casualties.

While our mathematical model accounts for multiple parameters and elements of an MCI, most of these elements are considered deterministically. Parameters such as travel distance and loading and unloading time could vary for each service, and the effect of non-deterministic values on the results needs to be studied. In addition, there could be cases in which exact number of the casualties is not available initially. We can study the case that number of the casualties increase in time as more casualties are triaged. Another area that needs more in-depth analysis is how casualty types vary over time as casualties condition degrades. Essentially, while a casualty is waiting for service, its survival probability decreases over time, and after some point, the casualty is practically in a more critical category. This dynamic casualty switching resembles the real world scenarios more precisely and could alter the structure of the optimal solution, thus, it needs a deeper analysis.

Chapter 5

Conclusion

Throughout this research, we show incorporating resources and scale of the disaster into the triage process can significantly improve the outcome in terms of total number of casualties survived. Using the developed models, we analyze the structure of the optimal policies under various settings and assumptions. One the more important assumptions is regarding how casualties' death is addressed in the modeling process. To the best of our knowledge, all the studies in the literature either assume all the casualties are served (no on-site death) or only the ones that are going to survived are served (no off-site deaths). While such simplifications allow deeper insight into the structure of the optimal solutions, they reduce the accuracy of the model. We develop a generalized model that can address both types of deaths. We analyze how such a generalization affects the results, and compare them with those of other models in the literature. In addition, we provide optimal algorithms under several settings, which can provide the optimal solution without solving any models. We show that these algorithms perform well under general setting, and thus, they can be used as high-quality heuristics.

To study all the steps of the triage process together rather than in isolation, we develop an integrated model including casualties prioritization, transportation, hospital assignment, and capacity management. This allows us to study how incorporation of several factors

affects the results. In practice several decisions need to be made simultaneously and the developed model could be a base for an emergency management decision support system. The developed models have the flexibility to address a broad range of circumstances containing multiple types of resources in different quantities. I study the trade-offs when there are multiple resources available and when availability of the resources gradually increase. I also focus on pediatric casualties due to their specific requirements to analyze how their inclusion affects the results and what the best strategies are in dealing with pediatric casualties.

The developed models in this research require few parameters following an MCI, including number of the casualties and the disaster location. Most of the other parameters can be either obtained or estimated with a high accuracy in advance. Many of the involved organizations often participate in emergency drills, simulating various types of incidents. Using such events emergency managers can gather data regarding hospital capacities and emergency vehicles available in their region. Having all the parameters, emergency personnel can generate an optimal policy for service and resource allocation. In future, we plan to collaborate with several emergency organizations to capture more details about the operations and make the model more practical and ready to be used. Another areas of future research include reverse triage, addressing multiple casualty locations, and self-triage. Reverse triage refers to early discharge of patients when resources at care facilities are overwhelmed. Considering reverse triage allows us to manage the limited hospital capacity more effectively. Also, in many types of MCIs there could be more than one disaster location, and thus, it is important to address this case and analyze how it affects the results.

Bibliography

Antommaria, A.H.M., Powell, T., Miller, J.E., Christian, M.D., for Pediatric Emergency Mass Critical Care, T.F., et al. (2011). "Ethical issues in pediatric emergency mass critical care." *Pediatric Critical Care Medicine*, **12 (6)**, pp. S163–S168.

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

Argon, N.T., Ziya, S., and Winslow, J.E. (2011). "Triage in the aftermath of mass-casualty incidents." *Wiley Encyclopedia of Operations Research and Management Science*.

Arnold, J.L. (2002). "Disaster medicine in the 21st century: future hazards, vulnerabilities, and risk." *Prehospital and Disaster Medicine*, **17 (01)**, pp. 3–11.

Aylwin, C.J., König, T.C., Brennan, N.W., Shirley, P.J., Davies, G., Walsh, M.S., and Brohi, K. (2007). "Reduction in critical mortality in urban mass casualty incidents: analysis of triage, surge, and resource use after the London bombings on July 7, 2005." *The Lancet*, **368 (9554)**, pp. 2219–2225.

Bagust, A., Place, M., Posnett, J.W., et al. (1999). "Dynamics of bed use in accommodating emergency admissions: stochastic simulation model." *BmJ*, **319 (7203)**, pp. 155–158.

Barbera, J.A. and Macintyre, A.G. (2007). *Medical surge capacity and capability: a management system for integrating medical and health resources during large-scale emergencies*. US Department of Health and Human Services.

Barfield, W.D., Krug, S.E., Kanter, R.K., Gausche-Hill, M., Brantley, M.D., Chung, S., Kissoon, N., for Pediatric Emergency Mass Critical Care, T.F., et al. (2011). "Neonatal and pediatric regionalized systems in pediatric emergency mass critical care." *Pediatric Critical Care Medicine*, **12 (6)**, pp. S128–S134.

Bazaraa, M.S., Jarvis, J.J., and Sherali, H.D. (2011). *Linear programming and network flows*. John Wiley & Sons.

Benson, M., Koenig, K.L., and Schultz, C.H. (1996). "Disaster triage: START, then SAVEa new method of dynamic triage for victims of a catastrophic earthquake." *Prehospital and Disaster Medicine*, **11 (02)**, pp. 117–124.

Bonnett, C.J., Peery, B.N., Cantrill, S.V., Pons, P.T., Haukoos, J.S., McVaney, K.E., and Colwell, C.B. (2007). "Surge capacity: a proposed conceptual framework." *The American journal of emergency medicine*, **25 (3)**, pp. 297–306.

Castle, N. (2006). "Triage and transport decisions after mass casualty incidents: NICK CASTLE explains the triage systems used at mass casualty incidents and looks at some of the issues that affect decisions about transport priorities." *Emergency Nurse*, **14 (1)**, pp. 22–27.

Cone, D.C. and MacMillan, D.S. (2005). "Mass-casualty triage systems: a hint of science." *Academic Emergency Medicine*, **12 (8)**, pp. 739–741.

Darr, K. (2006). "Katrina: Lessons from the aftermath." *Hospital topics*, **84 (2)**, pp. 30–33.

Dean, M.D. and Nair, S.K. (2014). "Mass-casualty triage: Distribution of victims to multiple hospitals using the SAVE model." *European Journal of Operational Research*, **238 (1)**, pp. 363–373.

Frykberg, E.R. (2002). "Medical management of disasters and mass casualties from terrorist bombings: how can we cope?" *Journal of Trauma-Injury, Infection, and Critical Care*, **53 (2)**, pp. 201–212.

Frykberg, E.R. and Tepas III, J. (1988). "Terrorist bombings. Lessons learned from Belfast to Beirut." *Annals of Surgery*, **208 (5)**, p. 569.

Gausche-Hill, M. (2009). "Pediatric disaster preparedness: are we really prepared?" *Journal of Trauma and Acute Care Surgery*, **67 (2)**, pp. S73–S76.

Ginter, P.M., Wingate, M.S., Rucks, A.C., Vásconez, R.D., McCormick, L.C., Baldwin, S., and Fargason, C.A. (2006). "Creating a regional pediatric medical disaster preparedness network: imperative and issues." *Maternal and child health journal*, **10 (4)**, pp. 391–396.

Griffiths, J.L., BAN, M., Kirby, N.R., ASM, M., HRD, B.B., Waterson, J.A., et al. (2014). "Three years experience with forward-site mass casualty triage-, evacuation-, operating room-, ICU-, and radiography-enabled disaster vehicles: Development of usage strategies from drills and deployments." *American journal of disaster medicine*, **9 (4)**, pp. 273–285.

Hick, J.L., Barbera, J.A., and Kelen, G.D. (2009). "Refining surge capacity: conventional, contingency, and crisis capacity." *Disaster Medicine and Public Health Preparedness*, **3 (S1)**, pp. S59–S67.

Hick, J.L., Koenig, K.L., Barbisch, D., and Bey, T.A. (2008). "Surge capacity concepts for health care facilities: the CO-S-TR model for initial incident assessment." *Disaster medicine and public health preparedness*, **2 (S1)**, pp. S51–S57.

Hoot, N.R. and Aronsky, D. (2008). "Systematic review of emergency department crowding: causes, effects, and solutions." *Annals of emergency medicine*, **52 (2)**, pp. 126–136.

Hupert, N., Hollingsworth, E., and Xiong, W. (2007). "Is overtriage associated with increased mortality? Insights from a simulation model of mass casualty trauma care." *Disaster Medicine and Public Health Preparedness*, **1 (S1)**, pp. S14–S24.

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

Iserson, K.V. and Moskop, J.C. (2007). "Triage in medicine, part I: concept, history, and types." *Annals of Emergency Medicine*, **49 (3)**, pp. 275–281.

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

Jenkins, J.L., McCarthy, M.L., Sauer, L.M., Green, G.B., Stuart, S., Thomas, T.L., and Hsu, E.B. (2008). "Mass-casualty triage: time for an evidence-based approach." *Prehospital and disaster medicine*, **23 (01)**, pp. 3–8.

Jones, N., White, M.L., Tofil, N., Pickens, M., Youngblood, A., Zinkan, L., and Baker, M.D. (2014). "Randomized trial comparing two mass casualty triage systems (JumpSTART versus SALT) in a pediatric simulated mass casualty event." *Prehospital Emergency Care*, **18 (3)**, pp. 417–423.

Kahn, C., Lerner, E., and Cone, D. (2010). "Triage." *Disaster Medicine: Comprehensive Principles and Practices. Cambridge University Press, New York, USA*, pp. 174–183.

Kamali, B. and Bish, D. (2016). "Service priority for mass-casualty incident response."

Kamali, B., Bish, D., and Glick, R. (2016). "Optimal service order for mass-casualty incident response."

Kanter, R.K. and Moran, J.R. (2007a). "Hospital emergency surge capacity: an empiric New York statewide study." *Annals of emergency medicine*, **50 (3)**, pp. 314–319.

Kanter, R.K. and Moran, J.R. (2007b). "Pediatric hospital and intensive care unit capacity in regional disasters: expanding capacity by altering standards of care." *Pediatrics*, **119 (1)**, pp. 94–100.

Kelen, G.D. and McCarthy, M.L. (2006). "The science of surge." *Academic Emergency Medicine*, **13 (11)**, pp. 1089–1094.

Kienstra, A.J. and Endom, E.E. (2002). "Bioterrorism and its impact on the emergency department." *Clinical Pediatric Emergency Medicine*, **3 (4)**, pp. 231–238.

Koenig, K.L., Cone, D.C., Burstein, J.L., and Camargo, C.A. (2006). "Surging to the right standard of care." *Academic Emergency Medicine*, **13 (2)**, pp. 195–198.

Kreis Jr, D.J., Fine, E.G., Gomez, G.A., Eckes, J., Whitwell, E., and Byers, P.M. (1988). "A prospective evaluation of field categorization of trauma patients." *Journal of Trauma and Acute Care Surgery*, **28 (7)**, pp. 995–1000.

Lerner, E.B., McKee, C.H., Cady, C.E., Cone, D.C., Colella, M.R., Cooper, A., Coule, P.L., Lairet, J.R., Liu, J.M., Pirrallo, R.G., et al. (2015a). "A Consensus-based Gold Standard for the Evaluation of Mass Casualty Triage Systems." *Prehospital Emergency Care*, **19 (2)**, pp. 267–271.

Lerner, E.B., Schwartz, R.B., Coule, P.L., Weinstein, E.S., Cone, D.C., Hunt, R.C., Sasser, S.M., Liu, J.M., Nudell, N.G., Wedmore, I.S., et al. (2008). "Mass casualty triage: an evaluation of the data and development of a proposed national guideline." *Disaster medicine and public health preparedness*, **2 (Supplement 1)**, pp. S25–S34.

Lerner, E.B., Schwartz, R.B., and McGovern, J.E. (2015b). "Prehospital triage for mass casualties." *Emergency Medical Services: Clinical Practice and Systems Oversight, Second Edition*, pp. 288–291.

Li, D. and Glazebrook, K.D. (2011). "A Bayesian approach to the triage problem with imperfect classification." *European Journal of Operational Research*, **215 (1)**, pp. 169–180.

Lyle, K., Thompson, T., and Graham, J. (2009). "Pediatric mass casualty: triage and planning for the prehospital provider." *Clinical Pediatric Emergency Medicine*, **10 (3)**, pp. 173–185.

Mace, S.E. and Bern, A.I. (2007). "Needs assessment: are Disaster Medical Assistance Teams up for the challenge of a pediatric disaster?" *The American journal of emergency medicine*, **25 (7)**, pp. 762–769.

Mills, A., Argon, N.T., and Ziya, S. (2011). "Resource-based START (ReSTART): mass-casualty triage under resource limitations." In *Annual Conference of Manufacturing & Service Operations Management*.

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

Nager, A.L. and Khanna, K. (2009). "Emergency department surge: models and practical implications." *Journal of Trauma and Acute Care Surgery*, **67 (2)**, pp. S96–S99.

Sacco, W.J., Navin, D.M., Fiedler, K.E., Waddell, I., Robert, K., Long, W.B., and Buckman, R.F. (2005). "Precise Formulation and Evidence-based Application of Resource-constrained Triage." *Academic emergency medicine*, **12 (8)**, pp. 759–770.

Sacco, W.J., Navin, D.M., Waddell, R.K., Fiedler, K.E., Long, W.B., Buckman Jr, R.F., et al. (2007). "A new resource-constrained triage method applied to victims of penetrating injury." *Journal of Trauma and Acute Care Surgery*, **63 (2)**, pp. 316–325.

Schneider, S.M., Gallery, M.E., Schafermeyer, R., and Zwemer, F.L. (2003). "Emergency department crowding: a point in time." *Annals of Emergency Medicine*, **42 (2)**, pp. 167–172.

Schrijver, A. (1998). *Theory of linear and integer programming*. John Wiley & Sons.

Sprivulis, P.C., Da Silva, J., Jacobs, I.G., Frazer, A.R., and Jelinek, G.A. (2006). "The association between hospital overcrowding and mortality among patients admitted via

Western Australian emergency departments." *Medical Journal of Australia*, **184 (5)**, p. 208.

Taylor, S., Jeng, J., Saffle, J.R., Sen, S., Greenhalgh, D.G., and Palmieri, T.L. (2014). "Redefining the outcomes to resources ratio for burn patient triage in a mass casualty." *Journal of burn care & research: official publication of the American Burn Association*, **35 (1)**, p. 41.

Toltzis, P., Soto-Campos, G., Kuhn, E., and Wetzel, R. (2013). "602: A Pediatric Triage Scheme to Guide Resource Allocation in a Mass Casualty." *Critical Care Medicine*, **41 (12)**, p. A148.

Toltzis, P., Soto-Campos, G., Kuhn, E.M., Hahn, R., Kanter, R.K., and Wetzel, R.C. (2015). "Evidence-Based Pediatric Outcome Predictors to Guide the Allocation of Critical Care Resources in a Mass Casualty Event*." *Pediatric Critical Care Medicine*, **16 (7)**, pp. e207–e216.

Wolf, P., Bigalke, M., Graf, B.M., Birkholz, T., and Dittmar, M.S. (2014). "Evaluation of a novel algorithm for primary mass casualty triage by paramedics in a physician manned EMS system: a dummy based trial." *Scand J Trauma Resusc Emerg Med*, **22 (1)**, p. 50.